Patel et al. BMC Genetics (2020) 21:84
https://doi.org/10.1186/s12863-020-00894-8

BMC Genetics

RESEARCH ARTICLE Open Access

CGGBP1-regulated cytosine methylation at CTCF-binding motifs resists stochasticity

Manthan Patel, Divyesh Patel, Subhamoy Datta and Umashankar Singh*

Abstract

Background: The human CGGBP1 binds to GC-rich regions and interspersed repeats, maintains homeostasis of stochastic cytosine methylation and determines DNA-binding of CTCF. Interdependence between regulation of cytosine methylation and CTCF occupancy by CGGBP1 remains unknown.

Results: By analyzing methylated DNA-sequencing data obtained from CGGBP1-depleted cells, we report that some transcription factor-binding sites, including CTCF, resist stochastic changes in cytosine methylation. By analysing CTCF-binding sites we show that cytosine methylation changes at CTCF motifs caused by CGGBP1 depletion resist stochastic changes. These CTCF-binding sites are positioned at locations where the spread of cytosine methylation *in cis* depends on the levels of CGGBP1.

Conclusion: Our findings suggest that CTCF occupancy and functions are determined by CGGBP1-regulated cytosine methylation patterns.

Keywords: CGGBP1, CTCF, Transcription factor binding sites, Cytosine methylation, Allelic imbalance, Stochasticity

Background

CTCF is a chromatin architectural protein with a broad repertoire of functions [1]. It is regarded as a principal regulator of higher-order chromatin structure. Maintenance of chromatin topology and chromatin boundaries are the key functions of CTCF [1–3]. The DNA-binding of CTCF is conventionally understood to take place through consensus DNA sequence motifs like M1 and M2 [4], Ren_20 [5] and LM2, LM7 and LM23 [5, 6]. Of the eleven Zn fingers in CTCF, one ZN7 interacts with cytosine in a methylation-sensitive manner [7]. This inhibition of CTCF-DNA binding and thus its function by motif methylation is a mechanism that regulates site-specific insulator activities of CTCF [8–12]. Methylation of CTCF-motifs and mitigation of CTCF function is a mechanism that has evolved to regulate the epigenome during development in a tissue-specific manner and has been reviewed extensively [13, 14]. While a lot of research has addressed the functions and regulatory effects of CTCF, the regulation of CTCF by partnering proteins has remained less studied. CTCF-interacting proteins such as YY1, the Cohesin complex and BRD2 for example, cooperate with CTCF and are needed for its enhancer-promoter looping, topological domains maintenance and boundary element functions respectively [15–19]. However, what regulates methylation at CTCF-motifs remains largely unknown. The regulator of methylation at CTCF motifs would naturally also be a regulator of CTCF-DNA binding.

Recently we have demonstrated that the human CGG triplet repeat binding protein CGGBP1 is required for normal genomic occupancy of CTCF [20]. CTCF not only binds to the DNA sequence-specific motifs, but also to interspersed repeats, mainly L1-LINEs and Alu-SINEs [20]. In the presence of CGGBP1, the repeat occupancy of CTCF accounts for more than 40% of all the binding sites, with L1 and Alu comprising the most of it. However, CGGBP1 depletion leads to an imbalance in the DNA-binding preference of CTCF. Upon CGGBP1 knockdown the repeat binding of CTCF is diminished and the CTCF-binding gets limited to the motifs [20]. Interestingly, like

* Correspondence: usingh@iitgn.ac.in
HoMeCell Lab, Biological Engineering, Indian Institute of Technology Gandhinagar, Palaj, Gandhinagar 382355, Gujarat, India

CTCF, CGGBP1 itself is a cytosine methylation-sensitive DNA-binding protein [20–25]. However, there is evidence that CGGBP1 binding to the target sequences prevents cytosine methylation from taking place [25, 26]. WGBS experiments have shown that CGGBP1 depletion leads to genome-wide disturbances in cytosine methylation [27, 28]. One of the major sites of cytosine methylation disturbances upon CGGBP1 knockdown are the Alu and L1 repeats. These sites double up as CGGBP1-binding sites as well [27, 29]. There seems to be an evolutionary relationship between the CGGBP1-binding SINEs and CTCF binding sites [4]. Thus the methylation regulation at Alu SINEs and CTCF-binding sites could thus have some hitherto unexplored evolutionary relationship as well. By maintaining the balance between CTCF occupancy at repeats or motifs, CGGBP1 acts as a regulator of CTCF binding pattern genome-wide. Data suggests that the repeat binding of CTCF takes place in cooperation with CGGBP1 [20]. Thus, while CGGBP1 depletion directly affects CTCF association with repetitive sequences, the gain of binding at motifs seems like an indirect consequence of CTCF displacement from the repeats. CGGBP1 however is also a methylation regulatory protein. Methylation changes at CTCF motifs can potentially affect CTCF binding and through it a change in the genome organization and function. We have previously shown that CGGBP1 depletion causes methylation disruption genome-wide with varied effects on repeats and sequence specific protein binding sites. These sites include regions that contain enhancers and known or predicted CTCF-binding sites [28]. CGGBP1-regulated methylation at CTCF-motifs could affect the binding of CTCF to motifs. The previous attempts to study the regulation of methylation by CGGBP1 using WGBS did not allow a high confidence detection of CTCF motifs in the sequence data [27, 28]. These studies in human fibroblasts revealed that CGGBP1 depletion causes a widespread disturbance in cytosine methylation. Gain as well as loss of methylation were observed at nearby cytosine residues genome-wide. Overall, CGGBP1 depletion resulted in a net increase in methylation levels.

Here we have used MeDIP-seq to analyze the cytosine methylation changes caused by CGGBP1 depletion in HEK293T cells or human skin fibroblast GM02639. We have used RNA interference against CGGBP1 (called the KD sample) and compared it against a non-targeting control (called the CT sample). We discover that there is a widespread disruption of methylation caused by CGGBP1 depletion in both the cell types with stochasticity being a major feature. Our results show that this stochasticity is partially explained by widespread allelic imbalances in cytosine methylation patterns between CT and KD. By a targeted analysis of transcription factor binding motifs from the JASPAR database, we report that CGGBP1 depletion disrupts methylation at a panel of potential transcription factor-binding sites (TFBSs). These TFBSs resist stochasticity and prominently include CTCF motifs. We have identified different cytosine methylation fates of CTCF-binding repeat-free motifs (RFMs) and motif-free repeats (MFRs) in CT and KD. Our analysis of MeDIP data in the flanks of RFMs show that these are CTCF-binding sites are required for maintenance of cytosine methylation patterns asymmetrically in the flanks of the CGGBP1-dependent CTCF-binding motifs. These findings provide evidence that cytosine methylation regulation by CGGBP1 can affect CTCF occupancy at RFMs with functional implications for cytosine methylation distribution *in cis*.

Methods

Cell culture, transfections and lentiviral transductions

HEK293T (NCCS, Pune), and human fibroblasts from Coriell Cell Repository (Son = GM02639, Parents = GM02641 and GM02640) were cultured in DMEM (HiMedia or HyClone) supplemented with 10% fetal bovine serum. HEK293T cells were subjected to lentiviral transduction with non-targeting shRNA (CT) and CGGBP1-targeting shRNA (KD) as described before [20]. The transduced cells were selected using Puromycin (10 μg/ml) for a week. The cells from three T75 flasks were lysed and genomic DNA was isolated using standard phenol:chloroform:isoamyl alcohol extraction method followed by ethanol precipitation and dissolution in 1xTris-EDTA buffer. For GM02639, the cells were transfected with non targeting siRNA (CT) or CGGBP1-targeting siRNA (KD) twice, once after 24 h and second after 72 h. The siRNA transfections were intended for a mild CGGBP1 depletion. The cells from the three T75 flasks were collected and processed for genomic DNA isolation as described above for HEK293T cells. Genomic DNA was isolated from the parental fibroblasts GM02641 and GM02640 without any transfections or transductions. The siRNA CGGBP1-targeting (4,392,422, ThermoFisher scientific) for KD and non-targeting (4,390,844, ThermoFisher scientific) for CT were transfected with the Oligofectamine™ Transfection Reagent (12,252,011, ThermoFisher Scientific). The transfections were performed as per the manufacturer's instructions.

Western blotting

The knockdown of CGGBP1 was confirmed by performing Western blots on the lysates of transduced HEK293T and transfected GM02639 cells at 96 h. The primary antibodies were a Rabbit anti-human CGGBP1 polyclonal (10716–1-AP, Proteintech) or Mouse anti-human GAPDH monoclonal (MA5–15738, Invitrogen). The secondary antibody was HRP conjugated Donkey anti-Rabbit (NA934, GE Life Sciences) or Sheep anti-Mouse (NA931, GE Life Sciences). The samples were resolved on 4–12% Bis-Tris NuPAGE (Invitrogen) gels, transferred to PVDF

membranes (3,010,040,001, Merck), blocked for 1 h in blocking buffer (5% dry milk w/v and fetal calf serum 10%v/v (HiMedia) in 1x TBST buffer) followed by overnight incubation with the primary antibody overnight at 4 °C (1:100 dilution in 1x TBST buffer). Membranes were washed in 1x TBST, incubated with HRP conjugated anti-rabbit secondary antibody (1:5000 dilution in blocking buffer) for 2 h at room temperature followed by washing with 1xTBST. The signals were developed using ECL substrate (Pierce) and captured using a chemiluminescence imaging setup (GE Life Sciences).

Methyl(cytosine) DNA immunoprecipitation (MeDIP)

The DNA isolated from HEK293T and GM02639 cells were sonicated to obtain DNA fragments of size range 150–300 bp. The conditions for sonication were 30 cycles of 30 s ON and 30 s OFF (Bioruptor, Diagenode) with intermediate mixing. 20 μg DNA for each sample was used as input for MeDIP. DNA was incubated with 1x MeDIP master mix (10 mM Sodium Phosphate Buffer, 0.14 M NaCl and 0.05% TritonX-100) containing a 5-methylcytosine antibody cocktail (5 μg; MABE146, SAB2702243; Sigma and NBP2–42813, Novus Biologicals) overnight with tumbling at 4 °C. Pre-washed Protein G sepharose beads (17–0618-01, GE Healthcare) were added to the mix and incubated for 2 h with tumbling mixing at 4 °C. The beads were allowed to settle, collected by gentle centrifugation and gently washed with 1x IP buffer three times. Further, the washed beads were incubated with 2 μl of 10 mg/ml Proteinase K (P2308, Sigma) in a protein digestion solution (50 mM Tris-HCl (pH 8.0), 10 mM EDTA (pH 8.0) and 0.5% SDS) containing for about 2 h at 56 °C with occasional mixing. The immunoprecipitated DNA was collected by subjecting the mix to centrifugation and collecting the supernatant into new tubes. The MeDIP DNA was purified using the PCR cleanup kit (A1460, Promega).

Sequencing of MeDIP and genomic DNA and quality control of sequencing output

The sequencing libraries for the MeDIP DNA (CT and KD from HEK293T and GM02639 cells) and genomic DNA (GM02641 and GM02640 cells) were generated according to the protocol mentioned elsewhere [20]. The sequencing was done using the Ion Proton S5 sequencer. The reads obtained after sequencing were filtered for poly-clonality and any PCR-duplications through the in-built default plug-in "FilterDuplicates"in the IonTorrent Suite.

Mapping and quality control

The reads obtained post-sequencing were controlled for the low quality reads through filtering out the reads having lower than q20 threshold. The initial QC was controlled through fastq validation using fastQValidator for filtration of any trimmed reads. The mean read length for all the samples was around 150 bp. The reads were mapped to repeat unmasked human genome hg38 with default mapping conditions using Bowtie2. Samtools was used for SAM to BAM conversions and sorting and indexing of the BAM file. Bedtools bedtobam and bamtobed functions were used for interconversions of BAM and BED files. The sequences from the reference genome (hg38 masked or unmasked) were obtained using bedtools getfasta option in bedtools. The fasta manipulations (format conversions, shuffle, statistics) were done using various functions in seqkit. The base composition and the cytosine contexts identification was done using the compseq function in the EMBOSS toolkit.

Variant calling

The mapped reads obtained as BAM output were subjected to variant calling using bcftools. The BAM file was subjected to mpileup followed by the bcftools call to identify variants across the sequenced locations for each dataset. The variants were filtered for their minimum coverage of 5 reads for each sample.

Methylation bin-wise read density distribution

To study the methylation spread, the methylation density was plotted in the methylation bins, with bin-size representing the number of methylation reads at the location. This bin-wise methylation read density was plotted for CT and KD of HEK29T and GM02639. Bins were categorised into low methylation bins (1 to 4), methylation bins 5 to 30 which account for differences between CT and KD and more than 30 methylation bins (not included in the bin-wise analysis).

MeDIP signal at 0.2 kb bins

Genome-wide 0.2 kb bins were created through bedtools make-windows option. The MeDIP signal for all the samples at these bins was obtained by bedtools coverage option with a minimum 50% of read overlapping with the bin. For pairwise analysis, the bins were retained with a minimum signal of three reads for CT and KD combined. The pairwise signal comparisons by Diff/Sum ((KD-CT)/(KD + CT)) was done for the signals obtained from HEK293T and GM02639 cells respectively. The frequency distribution of these Diff/Sum values across these filtered bins was plotted to compare the methylation changes for HEK293T and GM02639 cells respectively. The Diff/Sum ratios were calculated by normalizing the bam files to remove any artefacts due to unequal sequencing and alignment values between samples. The normalization values were a ratio of CT and KD aligned read counts and used to downsize the larger sample to the smaller sample. In addition, a randomized sampling of sequence reads was performed from the larger sample to match the read count

of the smaller sample and the Diff/Sum values were calculated. The random sampling of reads was done using bedtools sample function. The frequency distribution of these observed Diff/Sum values was compared for deviation (methylation disturbances leading to a net change in methylation) from an expected absolute normal distribution (methylation disturbances with no net change). The expected absolutely normal distribution was generated by artificially mirroring the positive and negative Diff/Sum values separately and taking mean values for each bin. The significance of deviation from normal was determined by the Kolmogorov-Smirnov test on bin-wise paired data.

Shannon entropy calculation

Briefly, stochasticity of DNA methylation is an unpredictable occurrence (ON state) or non-occurrence (OFF state) of methylation at a genomic location. This is due to a random binary choice between the ON and OFF states of methylation at any site [30]. The stochasticity in methylation changes upon CGGBP1 depletion in HEK293T and GM02639 cells was calculated as Shannon's entropy. The entropy was calculated as probability distributions of any 0.2 kb genomic bin (sequenced minimum 3 times in CT and KD combined, and at least 2 times in either sample) to be non-randomly different between GoM and LoM. For the range of differences which we used for Diff/Sum distributions, we plotted the random probabilities for the modulus of each Diff/Sum bin. For each of the bins the entropy values were calculated for the random probability of any region occurring in the state of no difference (|Diff/Sum| = 0) or difference (|Diff/Sum| = the bin value on X-axis) between CT and KD. An equal presence of a 0.2 kb region in GoM and LoM would give rise to an entropy value of 1 and a presence exclusively in GoM or LoM only would give rise to an entropy of zero.

JASPAR-wide motif identification

The overlapping coordinates between bed files were generated using bedtools intersect. The 0.2 kb bins were filtered for the threshold of |Diff/Sum| of more than 0.2 for HEK293T and GM02639 cells. These filtered bins were subjected to the motif identifications through the FIMO tool (MEME suite) for JASPAR-wide motifs (519 in total) using the stringent threshold of 1E-6. In parallel, the unfiltered bins genome-wide were also subjected to the same analysis to generate expected background motif occurrence frequencies. The randomisation of the bed coordinates was done using the bedtools shuffle option. These shuffled genomic coordinates were also subjected to motif finding using FIMO. The comparison of the expected and observed motif frequencies was performed for each transcription factor individually using Chi-square test function in Graphpad Prism 8.

Heatmap of MeDIP signal

The methylation changes across the 72 TFBSs were analysed and represented as a heatmap showing the extent of methylation change at these TFBSs in HEK293T and GM02639 cells. The Diff/Sum values between CT and KD at the bins representing each of the TFBSs was calculated for both cell types. The MeDIP signal was plotted on the bins for the filtered motifs as a Diff/Sum of average signal for CT and KD for each transcription factor. The plotting was done using the R package ComplexHeatmap tool using the single clustering method.

Heatmaps and average summary profile

The bigwig for the methylation signal was generated by using bamCoverage tools from deepTools. The scaling factor was applied to normalize the total readout of CT and KD samples for GM02639 cells. No scaling factor was applied for generation of bigwigs in CT and KD for HEK293T cells. The methylation signals were plotted as heatmaps using the deepTools plotHeatmap. The average summary plots along with standard deviation for methylation signals were plotted using plotProfile function with plotType std. option in deepTools. The matrix used for these functions were generated using deepTools computeMatrix function.

Correlation analysis

Correlation between MeDIP signals was performed by using the multiBigwigSummary tool from deepTools. Methylation signals were compared at bin sizes of 10 kb, 5 kb, 1 kb and 0.2 kb. Correlation between samples was calculated by Spearman method by using deepTools plotCorrelation tool.

PCA analysis

The PCA was performed for three sets of reads. These were the reads from RFM peaks, MFR peaks and reads genome-wide to compare the methylation signals between CT and KD of HEK293T and GM02639 cells. For PCA analysis, the reads obtained for the respective datasets were converted to bigwig format through bamCoverage function. These bigwigs were then subjected to multiBigwigSummary to obtain matrices. These matrices were used as inputs for PCA analysis through plotPCA function in deepTools. X-axes for all the three PCA plots represent the principal component 1 that accounts for the maximum variance among the four datasets. Y-axes for all the three PCA plots represent the principal component accounting for the second-largest variance among the four datasets.

Repeat content analyses

The repeat-masked and unmasked human genome (hg38) were used from the UCSC genome browser. Locally installed version (version open-4.0.9) of

RepeatMasker was used for repeat-masking. RMBlast (NCBI) was used as a repeat search engine and the repeat database used was Repbase (version available in 2018).

CTCF binding site identification and motif finding

The peaks were celled on the CTCF reads [20] that entirely overlapped with the methylated regions, across the methylation bins using MACS2.0 callpeak. De novo motif search was carried out by the locally installed version of MEME suite (version 5.0.3) tools meme. The motif search in GoM, LoM and no change 0.2 kb bins were performed by using default option with motif length 19 (−w value of 19). MEME suite tool FIMO was used to find predicted motifs in sequences. Predicted motifs were found with a default threshold of 1E-4.

Allelic imbalance in methylation

AIM was analysed for HEK293T and GM02639 upon CGGBP1 depletion. The minimum threshold of reads to analyse AIM was 5 for each sample. AIM was calculated as Diff/Sum for reads mapped to reference (Ref) and alternate (Alt) and was plotted for CT and KD for HEK293T as well as GM02639 cells. AIM along with its Parent-of-origin (PoO) for GM02639 cells was ascertained by pairing the loci common to all the four datasets (GM02639 CT and KD, GM02641 maternal and GM02640 paternal).

Statistical analyses, graphing and genome browser viewing

Statistical analysis was performed by using Prism version 8 (GraphPad) on numerical data stored and sorted in OpenOffice Spreadsheet. Visualisation of MeDIP signals at genomic regions were carried out by locally installed Integrated Genome Viewer.

Restriction digestion of genomic DNA for qPCR-based methylation detection

Genomic DNA was isolated from human dermal fibroblasts (106-05A,Sigma) transfected with Dharmacon siRNA cocktails as follows: Non-targeting (D-001910-10-20, Dharmacon) designated as CT and KD CGGBP1-targeting (E-015703-00-0020, Dharmacon) designated as KD. Genomic DNA was extracted as described above for MeDIP. DNA was sonicated to mean length of 1–1.5 kb and subjected to restriction digestion. CT and KD DNA were subjected to restriction digestion by methylation-dependent or methylation sensitive endonucleases. The digested DNA was used as templates for quantitative PCR for a panel of candidate regions (Additional file 1). The Ct values obtained from cytosine methylation (all contexts)-dependent digestion using *Mcr*BC were normalized against undigested input. Methylation sensitive digestions were performed using *Hpa*II and the Ct values were normalized using corresponding *Msp*I digestions. Following enzymes were used: *Hpa*II (R0171S, NEB), *Msp*I (R0106S, NEB) and *Mcr*BC (M0272S, NEB). For each restriction enzyme digestion, 1 μg of DNA template was used with 2 μl of enzyme for 6 h at 37 °C. The digested DNA was used as a template for qPCR using SYBR-Green PCR 2x mix (1,725,124, Bio-Rad), InstaQ96 (HiMedia) and the various primers as mentioned (Additional file 1). Relative quantitative changes were calculated using delta-delta Ct method.

Availability of sequence data

The MeDIP-seq data (HEK293T CT, HEK293T KD, GM02639 CT, GM02639 KD) and genomic DNA sequences (GM02640, GM02641) are publicly accessible at NCBI GEO vide GSE145300.

Results

CGGBP1 depletion causes widespread stochastic changes in cytosine methylation

The CT and KD HEK293T cells were generated as described elsewhere [20] (Additional file 2). DNA fragments were enriched using methylcytosine antibody and the methylated DNA (hereafter the DNA with cytosine methylation is referred to as methylated DNA) sequenced on the Ion Torrent platform with mean read lengths of 150 bp. The quality filtered deduplicated sequence reads were aligned to hg38. The alignment to hg38 was equally efficient in CT and KD (Additional file 3). Unlike the WGBS approach used earlier [27, 28], the MeDIP captured only the methylated DNA. Thus, the MeDIP-seq data expectedly did not show any significant differences in the sequence properties and base composition of CT and KD (Additional file 4). To characterize the differences in methylation between CT and KD, we compared the MeDIP signals (normalized read counts) for CT and KD in paired genomic bins (Fig. 1a). The correlation between CT and KD MeDIP signals varied strongly with the genomic bin size used for deriving the MeDIP signals. At 10 kb, the CT and KD methylation signals showed a near identity with a high Spearman correlation (Additional file 5). However, with a progressive decrease in the bin size down to 0.2 kb, this correlation was lost (Additional file 5). Randomization of CT and KD reads showed that the correlation at higher bin size and a loss of correlation at lower bin sizes is not due to random differences in CT and KD MeDIP (Additional file 6). Difference upon sum ratios (Diff/Sum) were calculated for normalized methylation signals in genomic bins of 0.2 kb paired between CT and KD (Fig. 1a). A frequency plot of the Diff/Sum values showed that there were large scale methylation disturbances genome-wide upon CGGBP1 depletion (Fig. 1b). Interestingly, as reported before [28], similar magnitude and frequency of gain of methylation

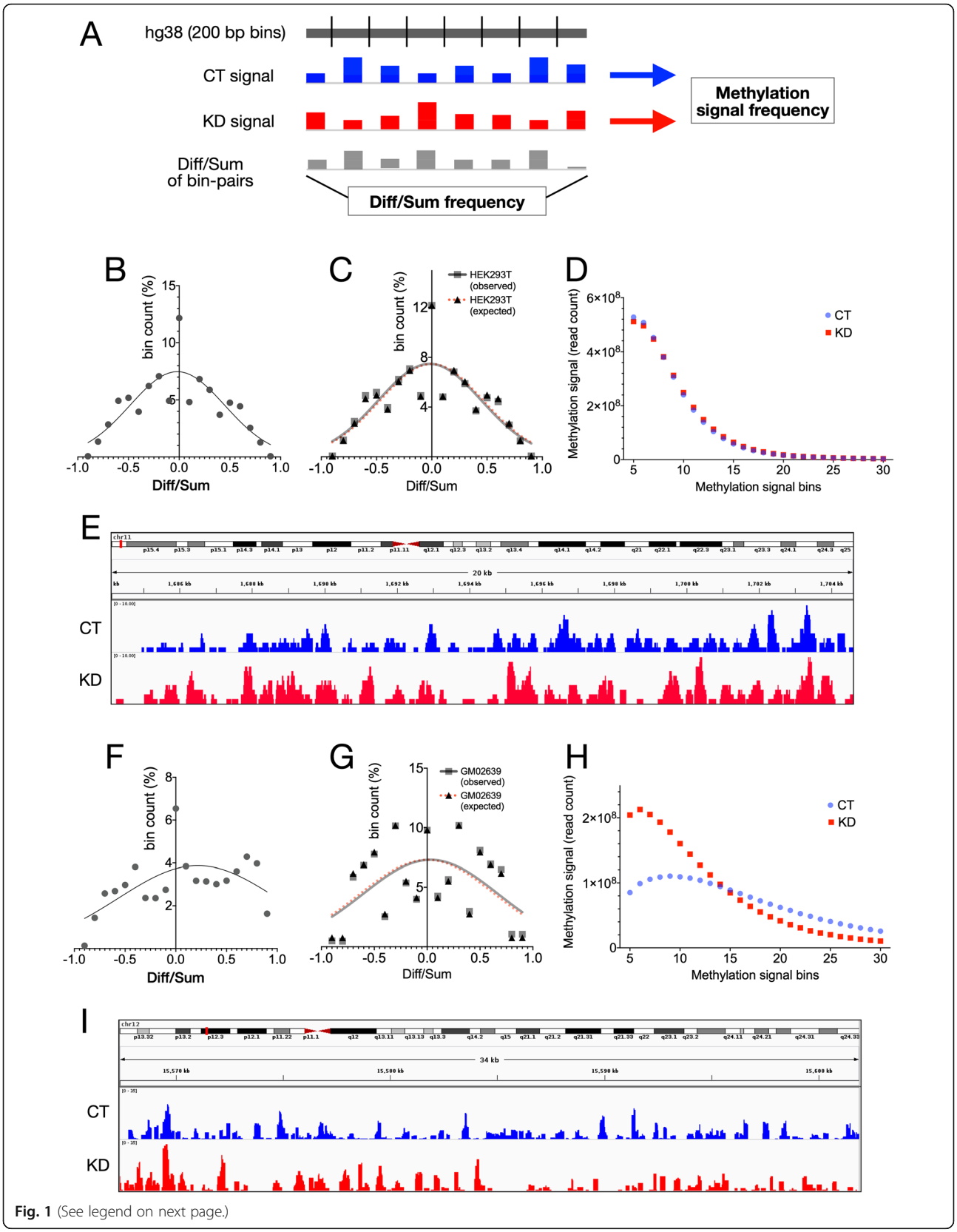

Fig. 1 (See legend on next page.)

(See figure on previous page.)
Fig. 1 Stochastic cytosine methylation patterns are selectively dependent on CGGBP1 depletion**a** A schematic representation of the two main methods used to quantify and compare MeDIP signals from CT and KD. The normalized MeDIP signal was obtained for CT (blue) and KD (red) in 0.2 kb bins of hg38 and the Diff/Sum ([KD-CT]/[KD + CT]) ratios were calculated for each bin pair. The frequency plots of Diff/Sum ratios and MeDIP signals were used to compare CT and KD. **b** Diff/Sum frequency in HEK293T shows a stochastic distribution resulting in a near-congruent gain-of-methylation (GoM) and loss-of-methylation (LoM). **c** A comparison of observed distribution and artificially generated expected distribution of methylation signal Diff/Sum ratios (see methods for details; plotted with a frequency interval of 0.1) for HEK293T shows a left-shift towards the negative Diff/Sum values (mean of differences − 0.2486; *p* value 0.0018). **d** The widespread GoM and LoM, as shown in **b,** nullify any net change in cytosine methylation resulting in highly similar methylation frequencies in HEK293T CT and KD. **e** Representative genome-view of HEK293T CT and KD MeDIP signals. **f** The Diff/Sum ratio distribution for GM02639 has a skewed distribution showing a net GoM. **g** Similar to the analysis in **c**, a comparison of observed and expected distributions for GM02639 showed a right-shift towards the positive Diff/Sum values (mean of differences 0.3609; p value 0.0009). **h** The MeDIP signal distribution for GM02639 CT and KD show that the GoM and LoM are restricted to regions with low and high methylation levels respectively. **i** A representative genome browser view of MeDIP signals for CT and KD.

(GoM) and loss of methylation (LoM) were observed (Fig. 1b).

For a quantitative assessment of the changes in methylation levels in HEK293T, we binned the methylation signal genome-wide into discrete units of signals ranging from a minimum of 5 to a maximum of 30 (genomic locations represented at least 5 times to a maximum of 30 times respectively in each MeDIP-seq data) (Fig. 1a). The rarely captured methylated DNA (signals bins less than 5) expectedly accounted for a large fraction of sequence reads in CT and KD (Additional file 7) whereas the signals diminished in bins more than 30 (not shown). A random sample of reads from CT matching the number of KD reads was used to compare observed and expected distributions of Diff/Sum ratios. The observed distribution showed a negative change with mean Diff/Sum ratio of − 0.2486 (Fig. 1c). Consistent with the Diff/Sum distribution (Fig. 1b), this methylation signal bin-wise distribution also revealed a near-identical distribution of methylation in CT and KD (Fig. 1d). A representative genome browser view showed that both gain and loss of methylation indeed occurred at nearby locations (Fig. 1e).

We extended the MeDIP analyses to primary fibroblast GM02639 to relate the above mentioned findings in HEK293T with the previously reported methylation regulation by CGGBP1 in foreskin fibroblasts. Using siRNA against CGGBP1, we transiently knocked down CGGBP1 (Additional file 8). In the previous studies, we have found the primary fibroblasts to be very sensitive to CGGBP1 depletion with a robust shutdown of transcription and exhibition of a stress-like phenotype [29]. To circumvent that, here we aimed at studying methylation changes caused by an incomplete depletion (approximately 50% knockdown by siRNA) of CGGBP1 in GM02639. The MeDIP-seq data from GM02639 CT and KD were normalized to eliminate any artefacts due to sequencing depth differences between the samples (Additional file 3). By a paired comparison of methylation in 0.2 kb bins genome-wide, we found that similar to HEK293T, the GM02639 also showed a widespread disturbance in methylation. However, there was a net increase in methylation in GM02639 KD compared to CT (Fig. 1f). For further scrutiny of this finding, we sub-sampled an equal number of reads from CT to match the count of KD and performed this analysis again. Since this random sub-sampling is expected to capture reads predominantly from the most prevalent low methylation bins and thus under-represent the methylation difference. The results corroborated the findings that unlike in the HEK293T cells, in GM02639 CGGBP1 depletion caused a net gain of methylation with a positive change of 0.3609 in Diff/Sum values (Fig. 1g). Similar to HEK293T, in GM02639 the correlation between CT and KD signals at 10 kb decreased drastically as we increased the methylation difference resolution to 0.2 kb (Additional file 9). A frequency plot of the number of genomic regions represented for a range of methylation signals (from 5 to 30) showed that the representation of weakly methylated regions was increased in KD (Fig. 1h). The majority of rarely captured methylation signals in bins 1 to 4 (Additional file 10) and the diminished population of reads in methylation bins > 30 (not shown) were excluded from this analysis. These results meant that the net increase in methylation was actually due to a much larger population of bins representing regions with low methylation signal in KD than in CT. A genome browser view of the representative positive as well as negative delta-signal regions showed that both gain and loss of methylation occurred at nearby locations (Fig. 1i). These findings were similar to the previous methylation analyses in fibroblasts where CGGBP1 depletion showed coincidental gain and loss of methylation with a marginal net gain of methylation [27].

The contents of satellite, Alu and LINE repeats were determined by using RepeatMasker on CT and KD datasets and no significant differences were found (not shown). However, when we plotted these repeat contents against methylation signal bins, we found that there was no net quantitative difference in methylation at repeats in HEK293T CT and KD (Additional file 11). In GM02639 only satellite repeats showed a significant methylation increase in KD (Additional file 12). These results were aligned with the previously reported findings that upon CGGBP1 depletion, the regions that carry high levels of

methylation undergo a further increase in methylation [27]. Interestingly, although the GoM and LoM in HEK293T stochastically cancelled out each other, the methylation change analysis at LINE and Alu repeat subfamilies revealed specific changes. Some subfamilies exhibited GoM and others showed LoM restricted to higher methylation bins (Additional file 11). The LINE and Alu repeats were also affected in GM02639 only at highly methylated regions (Additional file 12). These included stochastic changes in methylation upon CGGBP1 depletion and methylation change at L1 repeats and their subfamilies.

We could not identify any sequence motifs or related sequence properties that were different between the CT or KD MeDIP DNA. To objectively assess the stochasticity of methylation changes between CT and KD, we measured Shannon's entropy for the methylation changes (described in methods). The entropy analyses showed that the overall entropy was very high for CT and KD of HEK293T as well as the GM02639 cells indicating a high stochasticity in methylation states. The stochasticity was however non-uniformly distributed for the GM02639 cells (Fig. 2a). At higher magnitude of Diff/Sum ratios, the randomness was at its lowest suggesting that by applying a stringent cutoff for methylation change, we could extract the non-stochastic determinants of methylation change between CT and KD.

We thus applied a combination of three filters to extract and study deterministic changes in methylation: differentially methylated between CT and KD in HEK293T as well as GM02639, a minimum |Diff/Sum| value of 0.2, and a minimum sequence coverage of 3 reads per 0.2 kb bin for CT and KD combined. Using these criteria we asked if the 0.2 kb regions undergoing GoM or LoM are differentially enriched in DNA sequence motifs that constitute known protein binding sites.

CGGBP1 regulated methylation patterns target multiple TFBSs including those of CTCF

Methylation is a major regulator of DNA binding of transcription factors (TFs) [31, 32]. We asked if the methylation disturbance caused by CGGBP1 depletion affects known transcription factor binding sites (TFBSs).

A stringent search ($p < 1E\text{-}6$) for JASPAR motifs of 519 TFs was performed in 0.2 kb bins covered in the CT or KD MeDIP dataset with a minimum coverage of 3 reads. This analysis showed that highly probable binding sites for more than 300 TFBSs are immunoprecipitated in MeDIP DNA of CT as well as KD (Fig. 2b). In this search, the well-known chromatin regulator protein CTCF featured as one of the proteins with the highest occurrence in the CT and KD MeDIP DNA for both HEK293T and GM02639 (Fig. 2b, orange data point). This constituted the expected frequency of TFBS occurrence in the combined MeDIP datasets. Subsequently, the TFBS frequencies were calculated in those genomic bins where the normalized methylation signals were different between CT and KD (|Diff/Sum| > 0.2) (Fig. 2c, CTCF highlighted in orange). These observed TFBS frequencies for 343 TFs were compared against the expected frequencies and analyzed for each TF separately. A total of 72 TFs showed a significantly higher presence of TFBS in the observed (occurrence in bins differentially methylated between CT and KD) as compared to the expected. The methylation signal Diff/Sum ratios were calculated for these TFBS separately in HEK293T and GM02639 datasets (Fig. 2b and c). Interestingly, most of these 72 TFs, including CTCF, showed opposite changes in methylation in HEK293T and GM02639 (Fig. 2d). There were two major clusters which showed different patterns of methylation change in HEK293T (upper panel) and GM02639 (lower panel) upon CGGBP1 depletion (Fig. 2d). One of these clusters was of the TFBS undergoing GoM in GM02639 and LoM in HEK293T. Another major cluster was of protein with weaker methylation changes in either cell line. We pursued the first cluster further to study how despite a stochastic methylation change, a set of TFBSs continue to exhibit a directional change in methylation. One of the 29 TFs contained in this cluster was CTCF, which is known to bind to DNA in a methylation-sensitive manner and is also regulated by CGGBP1.

CTCF occupancy at motifs as compared to repeats depends on the levels of CGGBP1 as has been demonstrated in HEK293T cells [20]. Whether the changes in cytosine methylation caused by CGGBP1 depletion play a role in determining CTCF binding to its motifs or its occupancy at repeats is not known. As a step towards understanding this possibility, we analyzed the methylation changes at CGGBP1-dependent CTCF-binding sites.

The nature of CTCF-binding DNA sequence motifs is different between CT and KD with a G/C weightage difference at position eight [20]. However, the CTCF motifs present in HEK293T GoM and LoM fractions (see Fig. 1b) showed no such difference and resembled the canonical CTCF motif (Additional file 13). The ChIP-seq reads which were pulled down in KD represented the motif-rich regions of the genome which remain bound to CTCF in the absence of CGGBP1 [20]. These regions, although motif-rich, are expected to maintain low methylation levels as compared to CT. We fetched these HEK293T CT and KD reads from the published CTCF ChIP-seq data [20] and measured methylation signals at them. As expected, the CTCF-bound CT and KD reads were distributed in CT and KD MeDIP data with the same pattern as their distribution curves across methylation bins were overlapping. However, both the CT and KD curves showed a concentration near low methylation bins (Fig. 3a). These reads also gave rise to

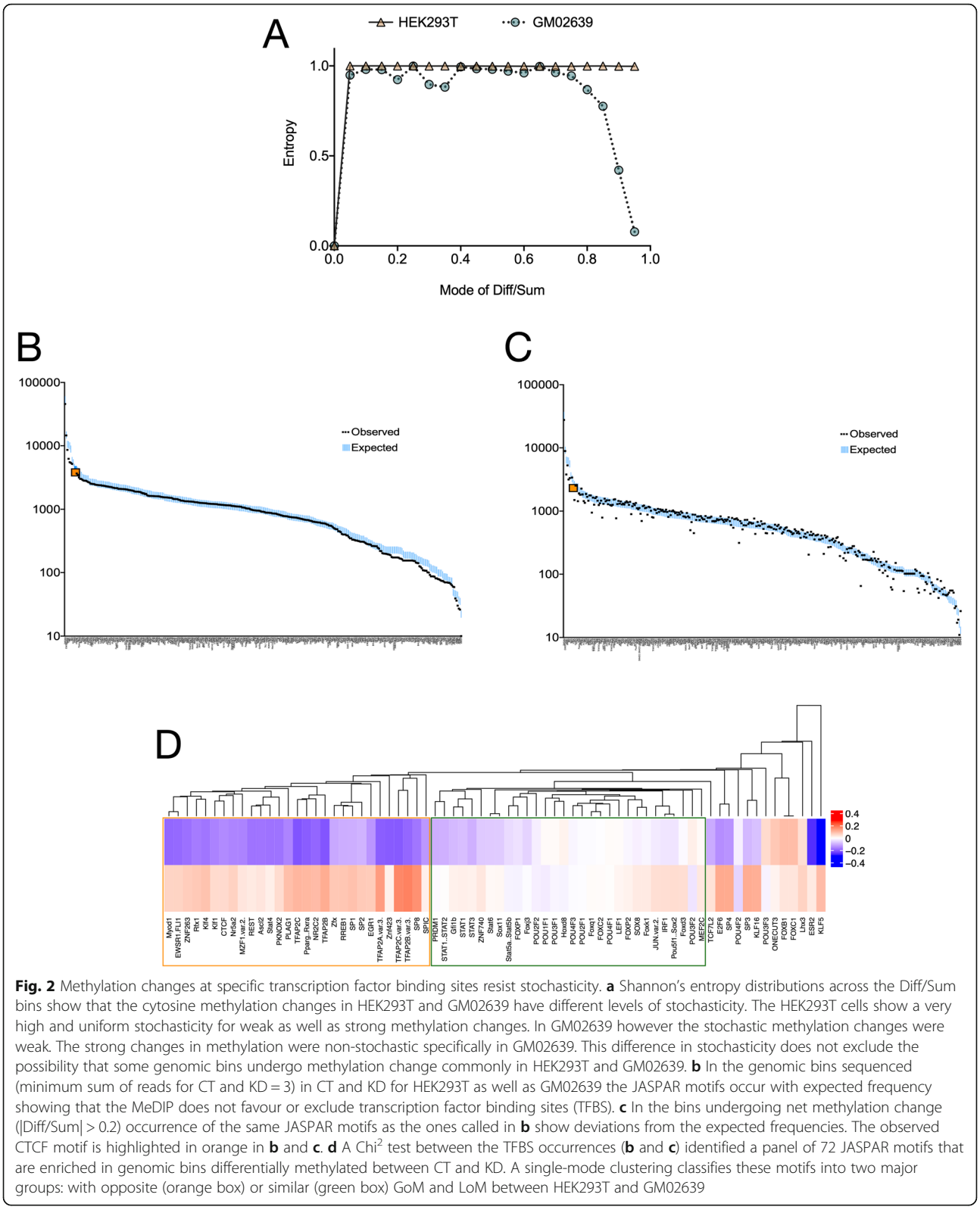

Fig. 2 Methylation changes at specific transcription factor binding sites resist stochasticity. **a** Shannon's entropy distributions across the Diff/Sum bins show that the cytosine methylation changes in HEK293T and GM02639 have different levels of stochasticity. The HEK293T cells show a very high and uniform stochasticity for weak as well as strong methylation changes. In GM02639 however the stochastic methylation changes were weak. The strong changes in methylation were non-stochastic specifically in GM02639. This difference in stochasticity does not exclude the possibility that some genomic bins undergo methylation change commonly in HEK293T and GM02639. **b** In the genomic bins sequenced (minimum sum of reads for CT and KD = 3) in CT and KD for HEK293T as well as GM02639 the JASPAR motifs occur with expected frequency showing that the MeDIP does not favour or exclude transcription factor binding sites (TFBS). **c** In the bins undergoing net methylation change (|Diff/Sum| > 0.2) occurrence of the same JASPAR motifs as the ones called in **b** show deviations from the expected frequencies. The observed CTCF motif is highlighted in orange in **b** and **c**. **d** A Chi^2 test between the TFBS occurrences (**b** and **c**) identified a panel of 72 JASPAR motifs that are enriched in genomic bins differentially methylated between CT and KD. A single-mode clustering classifies these motifs into two major groups: with opposite (orange box) or similar (green box) GoM and LoM between HEK293T and GM02639

genuine peaks with CTCF-motifs, the distributions of which across the methylation bins also followed the same pattern as those of the reads and showed high peak presence at low methylation bins (Fig. 3b).

The findings with the JASPAR-wide motif search (Fig. 2c) showed that the effect of CGGBP1 depletion on CTCF motifs in GM02639 would be different from that observed in HEK293T. The MeDIP signals were then used to identify how the methylation patterns were affected in GM02639 at the CTCF-bound DNA in CT and KD. However, the GM02639 methylation bins also corresponded to CTCF ChIP-seq reads in a way similar to HEK293T as the low methylation bins were rich in reads. Strikingly, the identification of genuine peaks and motifs in those peaks was also restricted to low methylation bins in KD as compared to CT (Fig. 3c and d). These similarities supported the assumption underlying this analysis that common mechanisms operate to regulate CTCF-binding to motifs in HEK293T and GM02639 and that these common mechanisms may involve methylation regulation by CGGBP1. However, the enrichment of CTCF-bound reads at low methylation bins was different between CT and KD only in GM02639, not HEK293T. Thus, despite the indistinguishable methylation sensitivity of CTCF-binding to motifs in HEK293T and GM02639, their dependencies on CGGBP1-regulated methylation were different in the two cell lines. Unlike HEK293T, in GM02639 the KD read and peak distribution curves (Fig. 3c and d respectively) crossed the CT curve demonstrating that the methylation sensitivity of CTCF-motif binding is non-stochastically higher in KD than CT.

A panel of demonstrated and possible CTCF-binding sites were selected for analyzing methylation changes caused by CGGBP1 depletion to validate the finding that methylation at CTCF-binding sites is regulated by CGGBP1. By using qPCR on fibroblast CT and KD genomic DNA digested with methylation-sensitive or methylation-dependent restriction endonucleases, we established that methylation levels at many CTCF-binding sites are indeed affected by CGGBP1 depletion (Additional file 14). The MeDIP reads at CTCF-motifs could be captured in the assay due to methylation at cytosine residues outside the CTCF-motifs. However, even if the MeDIP enrichment were to be associated with the methylation change within the CTCF motifs, a combined analysis of the motifs identified in MeDIP reads at CTCF-motifs and the qPCR results suggested that these were non-CpG methylation events.

We then cross-validated these findings by analyzing the CT and KD methylation signals at the previously characterized CGGBP1-regulated CTCF-binding sites. The disturbances observed in methylation patterns at CT peaks (Fig. 3e) or KD peaks (Fig. 3f) for HEK293T were weaker than the same observed for GM02639 CT (Fig. 3g) or KD peaks (Fig. 3h; compare the grey signals in Fig. 3e and Fig. 3f with Fig. 3g and Fig. 3h respectively). Noticeably, in GM02639, where methylation at CTCF-binding sites and flanks were increased by CGGBP1 depletion (Fig. 3c and d), the binding of CTCF was retained and restricted in KD to regions with much lower methylation levels than CT (Fig. 3g compared with Fig. 3h). Overall, the occurrence of CTCF-binding sites in GoM or LoM regions as well as occurrence of GoM and LoM at CTCF-binding sites together established that CGGBP1 depletion causes targeted methylation changes at CTCF-binding sites and its flanks in a cell type-specific manner. However, CTCF-binding sites at repeats and motifs show opposite changes in CTCF occupancy upon CGGBP1 depletion. If methylation regulation by CGGBP1 is a potential means to regulate CTCF binding, then CGGBP1 depletion would cause different patterns of methylation changes at CTCF-binding repeats and motifs. We thus performed a comparative analysis of methylation change at CTCF-binding repeats and motifs.

CGGBP1 affects methylation at CTCF-binding repeat-free motifs and motif-free repeats differently

We have described that the MFR and RFM constitute exclusive sets CGGBP1-regulated CTCF-binding sites [20]. Whereas CTCF occupancy at MFR depends on CGGBP1, the same at RFM is not clearly understood. As described above, methylation changes caused by CGGBP1 depletion at CTCF-binding sites are concentrated at motifs with no specific changes at the repeats. To test this more rigorously, we focussed on MFR and RFM for a comparative analysis of methylation changes caused by CGGBP1 depletion at CTCF-binding sites.

The methylation signals at RFM and MFR were compared and Diff/Sum values calculated for paired bins between CT and KD. Methylation disturbances were normally distributed in HEK293T at MFR as well as RFM. Unlike MFR, the Diff/Sum value distribution in RFM showed a slight positive skew (Fig. 3i), which was in agreement with the findings of methylation change at CTCF motifs described in Fig. 3a. In GM02639, the Diff/Sum distribution of methylation changes at MFR were normally distributed with an approximately 30% reduction in the count of bins showing no methylation change (Fig. 3j) compared to that in HEK293T (Fig. 3i). The fraction of RFM undergoing a methylation change with $|\text{Diff/Sum}| = 0.5$ was more than two folds higher in GM02639 than in HEK293T. The Diff/Sum value distribution of RFM in GM02639, however, showed a clear deviation from a normal distribution with three modes. It also showed a marked increase in the non-zero Diff/Sum population, which represents the fraction undergoing methylation change (Fig. 3j). These results were consistent with the findings that methylation changes in

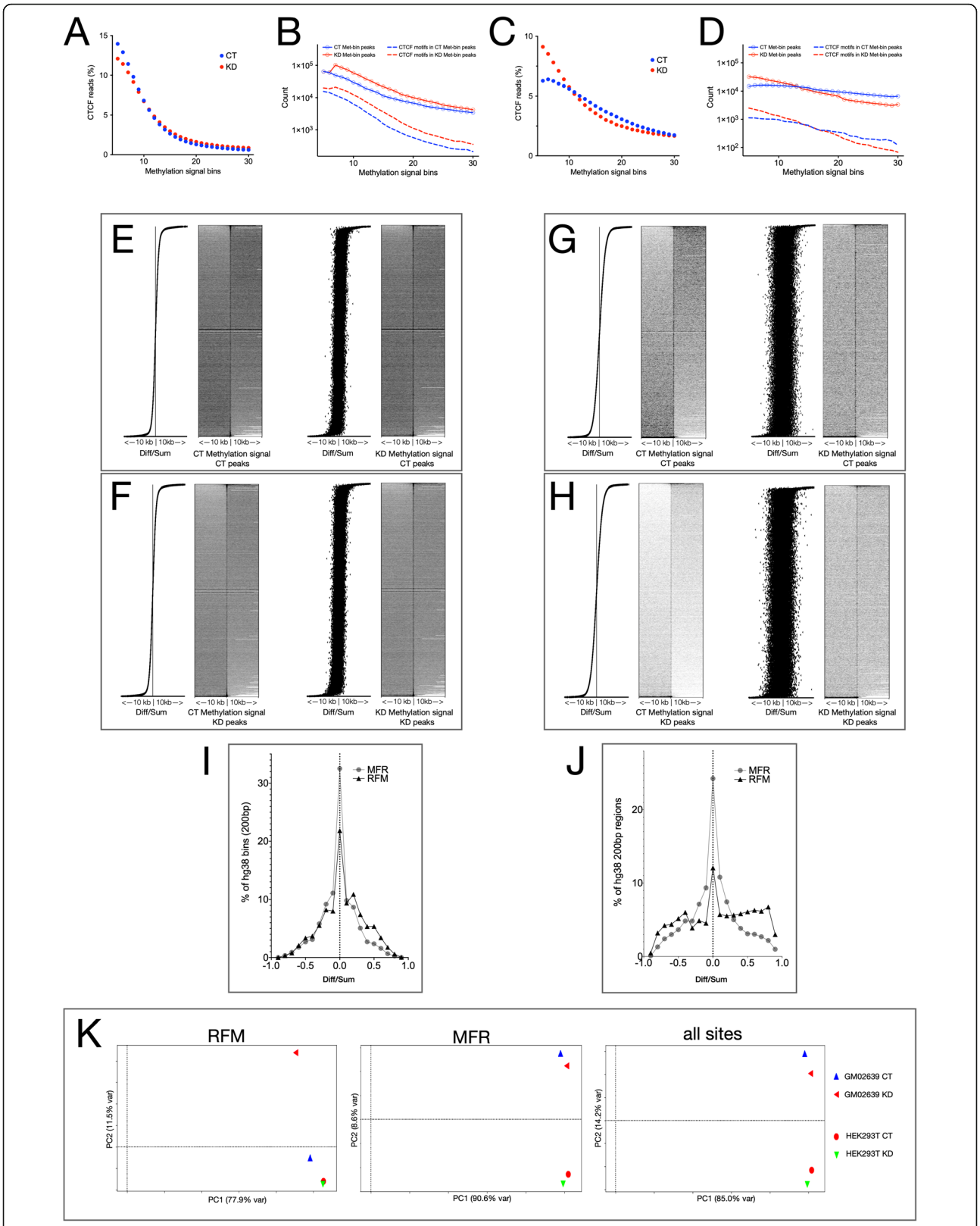

Fig. 3 (See legend on next page.)

(See figure on previous page.)
Fig. 3 Cytosine methylation changes caused by CGGBP1 depletion are less stochastic at CTCF-binding RFM than at MFR. **a** CTCF-binding sites in HEK293T are enriched in low methylation bins with no strong differences in MeDIP signals between CT and KD. **b** MeDIP reads in HEK293T CT and KD are equally enriched in peaks (dashed lines) that are positive for CTCF motifs (continuous lines-circles). In agreement with the cytosine methylation-sensitivity of CTCF binding to DNA, the occurrence of peaks and CTCF motifs decline with an increase in MeDIP signal. **c** In GM02639 the concentration of CTCF-binding sites at low methylation bins is enhanced by CGGBP1 depletion. **d** The occurrence of peaks in the GM02639 MeDIP reads (dashed lines) and their CTCF motif positivity (continuous lines-circles) regresses at a higher rate in KD showing that the cytosine methylation sensitivity of CTCF-binding at motifs in GM02639 is stronger in absence of CGGBP1. **e** and **f** Cytosine methylation patterns at CTCF-binding sites in CT (**e**, $n = 42{,}978$ CTCF peaks) or KD (**f**, $n = 47{,}632$ CTCF peaks) are disrupted upon CGGBP1 depletion as revealed by MeDIP signals in the 10 kb flanks of the peak centres (X-axes). **g** and **h** GM02639 MeDIP signals in the 10 kb flanks of peak centres (X-axes) at the same CTCF-binding sites (as shown in **e** and **f**) highlight three differences from the pattern observed in HEK293T (**g** and **h** compared with **e** and **f** respectively); overall low cytosine methylation levels, a stronger loss of methylation in KD compared to CT, and a larger disturbance in methylation caused by CGGBP1 depletion. A comparison of **e** and **f** with **g** and **h** respectively reinforces the findings that the cytosine methylation difference between CT and KD is less stochastic (Fig. **2a**) and the CTCF-binding motifs in KD are recused to a low methylation status in GM02639 (Fig. 2**c** and **d**, and Fig. 3**c** and **d**). **i** The cytosine methylation changes observed in HEK293T (**a**, **b**, **e** and **f**) affect the repeat-derived (MFR) and motif-derived (RFM) CTCF-binding sites differently with a slight net GoM at RFM. **j** The cytosine methylation changes in GM02639 CT and KD were strongly different between MFR (stochastic GoM and LoM with a normal distribution of Diff/Sum) and RFM (strong GoM and LoM with a multimodal distribution of Diff/Sum). A comparison of **i** and **j** shows that the cytosine methylation at RFM, unlike MFR, is specifically regulated by CGGBP1. The CGGBP1-dependence of RFM cytosine methylation is higher in GM02639 than HEK293T. **k** PCA analyses reveal the different levels of stochasticity of methylation changes at RFM and MFR. The major component of variance (PC1) shown on the X-axis represented stochastic changes as it failed to segregate the CT and KD samples of the two cell types when RFM and MFR were analyzed separately or together. Commensurate with the previously described findings, the PC1 (X-axis) accounted for the least stochastic variance in RFM (77.9%) and highest (90.6%) for MFR. The PC2 (Y-axis) accounted for 11.2% of variance between GM02639 CT and KD RFM only. For MFR and all sites, the PC2 (Y-axis) accounted for variances between the cell types but not CT versus KD. Thus, at RFM the MeDIP signals have a GM02639-specific dependence on CGGBP1 whereas the same at MFR follow a cell type-specific pattern predominantly

GM02639 due to CGGBP1 depletion were more pronounced than those in HEK293T.

The Diff/Sum distributions of stochastic changes in methylation are expected to be normal. The deviations from a normal distribution indicate a specific association between RFM and methylation change in KD as compared to CT. The net methylation changes are however a sum of stochastic changes and specific changes. We performed PCA analysis to find out how the RFM and MFR methylation changes in CT and KD are different between HEK293T and GM02639.

As shown in Fig. 3k, the largest principal component that accounted for most of the variance (the X-axes) failed to differentiate either the two cell lines or the samples CT and KD. The percentage of variance accounted for by this major component was 78% at RFM and 90% at MFR and 85% when all methylation changes across the genome (hg38) were measured (Fig. 3k). This large component of variance not accounting for differences between the samples reflected the stochasticity of methylation changes. The second principal component (the Y-axes in Fig. 3k) accounted for the variation between GM02639 CT and KD at RFM only. For MFR and hg38, the second principal component only accounted for differences between the two cell lines. Thus, the difference between the methylation patterns at CT RFM and KD RFM in GM02639 was the only non-stochastic change in methylation across all the combinations of RFM, MFR and CT or KD in the two cell lines. Down to the fourth principal component (accounting for > 99.99% of variance) the CT and KD could not be differentiated at MFR in either cell line (not shown).

These analyses confirmed that CGGBP1 regulates methylation at RFM in GM02639 specifically. We argued that specific regulation of methylation at RFM in GM02639 should also manifest as a non-stochastic and predictable pattern of methylation change between CT and KD at RFM and not MFR. To pursue this, we compared the methylation patterns in the flanking regions of the RFM and MFR.

CTCF-binding RFM correspond to methylation transition boundaries

CTCF binding at the MFR has been shown to be required for restriction of H3K9me3 spread. Ablation of CTCF binding at repeats results in a disruption in H3K9me3 levels in the flanks of the MFR with most MFR exhibiting a loss of barrier activity (LoB) upon CGGBP1 depletion. The current findings, that unlike MFR, the RFM are specifically associated with cytosine methylation changes, suggested that similar to the barrier activities of MFR for H3K9me3 levels, the RFM could act as barriers for cytosine methylation levels. The difference between upstream and downstream methylation signals in 10 kb flanks of RFM and MFR was calculated for CT and KD HEK293T. There was a widespread asymmetry in the methylation signals obtained from upstream and downstream flanks of RFM (Fig. 4a). The methylation level asymmetries in RFM flanks were poorly correlated between CT and KD (Fig. 4b). On the other hand, the asymmetries between methylation signals in the upstream and downstream flanks of MFR were higher than those observed for RFM (Fig. 4c; compared with Fig. 4a), yet highly correlated between CT

and KD (Fig. 4d). These findings in MFR were commensurate with a more stochastic change in methylation in MFR flanks as compared to RFM. Visualization of methylation signals in HEK293T (Fig. 4, e to h) showed that at some RFM a lower downstream (Fig. 4e) or upstream (Fig. 4f) methylation in CT is increased to yield a no asymmetry in KD. Similarly, a gain of methylation selectively upstream (Fig. 4g) or downstream (Fig. 4h) of CTCF-binding sites was observed at other RFM. This methylation level asymmetry in the flanks of RFM was expected to be more widespread in GM02639. Indeed a visualization of methylation signals in GM02639 CT and KD showed that a larger number of RFM showed a loss of methylation asymmetry due to an increase of methylation in the downstream (Fig. 4i) or upstream (Fig. 4j) flanks in KD. An even larger number of RFM showed a gain of methylation asymmetry due to an increase in methylation selectively in the upstream (Fig. 4k) or downstream (Fig. 4l) flanks in KD. These predictable and deterministic methylation changes occurring selectively at RFM could be of functional consequence for CTCF-binding to motifs and regulation of chromatin barrier activities. Interestingly, the RFM regions with cytosine methylation asymmetries were different from the repeat-rich CTCF-binding sites that act as barriers for H3K9me3 levels [20] and we could not find any overlap between them (not shown).

These sites for specific methylation regulation by CGGBP1, however, were embedded in a much larger fraction of the genomic regions at which methylation changes were stochastic. One possibility that could explain this stochastic methylation disruption is that upon CGGBP1 depletion the two allelic copies become amenable to methylation changes independently. Allele-specificity of CTCF-binding and its regulation by allele-specific methylation is an established mechanism. We wanted to find out if unexpected levels of allelic imbalance underlie the stochasticity in methylation changes caused by CGGBP1 depletion. Thus, we analysed the allelic imbalance in methylation and its occurrence in CT and KD MeDIP datasets.

Unexpected levels of allelic imbalance in MeDIP DNA upon CGGBP1 depletion

To find out the contribution of allelic imbalances towards stochasticity of methylation changes observed upon CGGBP1 depletion, we studied the proportions of alleles represented in MeDIP data separately for HEK293T and GM02639.

Allelic counts were obtained in HEK293T input [20] and compared with MeDIP CT and KD for all loci where both or either sample are heterozygous. The homozygosity or heterozygosity was called only if the locus was covered minimum five times in each sample. The presence of reference (Ref) and alternate (Alt) alleles constituted a heterozygous genotype, whereas the occurrence of only Ref or Alt was called homozygous. If CGGBP1 depletion did not affect methylation with an allelic bias, then the Alt and Ref genotypes would be represented equally in CT and KD and the overall genotype distributions for CT and KD MeDIP DNA would resemble that of the input. The input for HEK293T showed an expected skewed distribution with a higher presence of the Ref allele as compared to the Alt allele. The CT and KD MeDIP, however, showed a multimodal distribution with an unexpectedly high representation of the Ref as well as the Alt alleles (Fig. 5a). The observed allelic distributions in CT and KD were both disturbed and different from the expected distribution of alleles as seen in the input (Fig. 5b and c), but were highly similar to each other (Fig. 5d). Thus, the CT and KD MeDIP DNA had an unexpected genotype distribution clearly demonstrating an allelic imbalance. Also, the near congruence of the allelic ratio distributions showed that the deviation from the expected genotype ratios was stochastic.

The Ref and Alt alleles were called in GM02639 MeDIP data as well and the Ref and Alt genotype ratios for CT were plotted against those obtained from KD. We found that there was large scale allelic imbalance represented in the genotype distributions (Fig. 5e). The CT and KD genotype ratios were more anticorrelated in GM02639 ($r^2 = 0.000169$) as compared to that in HEK293T ($r^2 = 0.130321$).

To objectively establish the extent of allelic imbalance in MeDIP between CT and KD, we sequenced the paternal and maternal DNA for GM02639 (see methods for details). We fished out only those regions at which the two parents were homozygous for different alleles such that at these loci the GM02639 could only be heterozygous in the absence of any sweeping allelic imbalances in methylation. Out of 11,526 such loci, GM02639 was expected to be heterozygous with Mat^{Alt}/Pat^{Ref} genotype for 6613 loci and Mat^{Ref}/Pat^{Alt} genotype at 4913 loci. We set an arbitrary threshold of Diff/Sum ratio such that values ranging between − 0.5 and 0.5 were regarded as heterozygous (green shaded region of the scatter; Fig. 5f). In this case, the heterozygosity represented biallelic methylation within the range of Diff/Sum ratio threshold. Four types of unexpected deviations from the expected heterozygosity were observed in both CT and KD (non-green shaded regions of the scatter; Fig. 5f). These were as follows: Mat/− or −/Pat (due to a monoallelic methylation bias similarly in CT and KD; red shade in Fig. 5f), Mat/− or −/Pat in CT and Mat/Pat in KD (due to a loss of monoallelic methylation bias KD; aqua shade in Fig. 5f), Mat/− or −/Pat in KD and Mat/Pat in CT (due to a gain of monoallelic methylation bias KD; purple shade in Fig. 5f), and an allelic flip from Mat/− in CT to −/Pat in KD and from −/Pat in CT

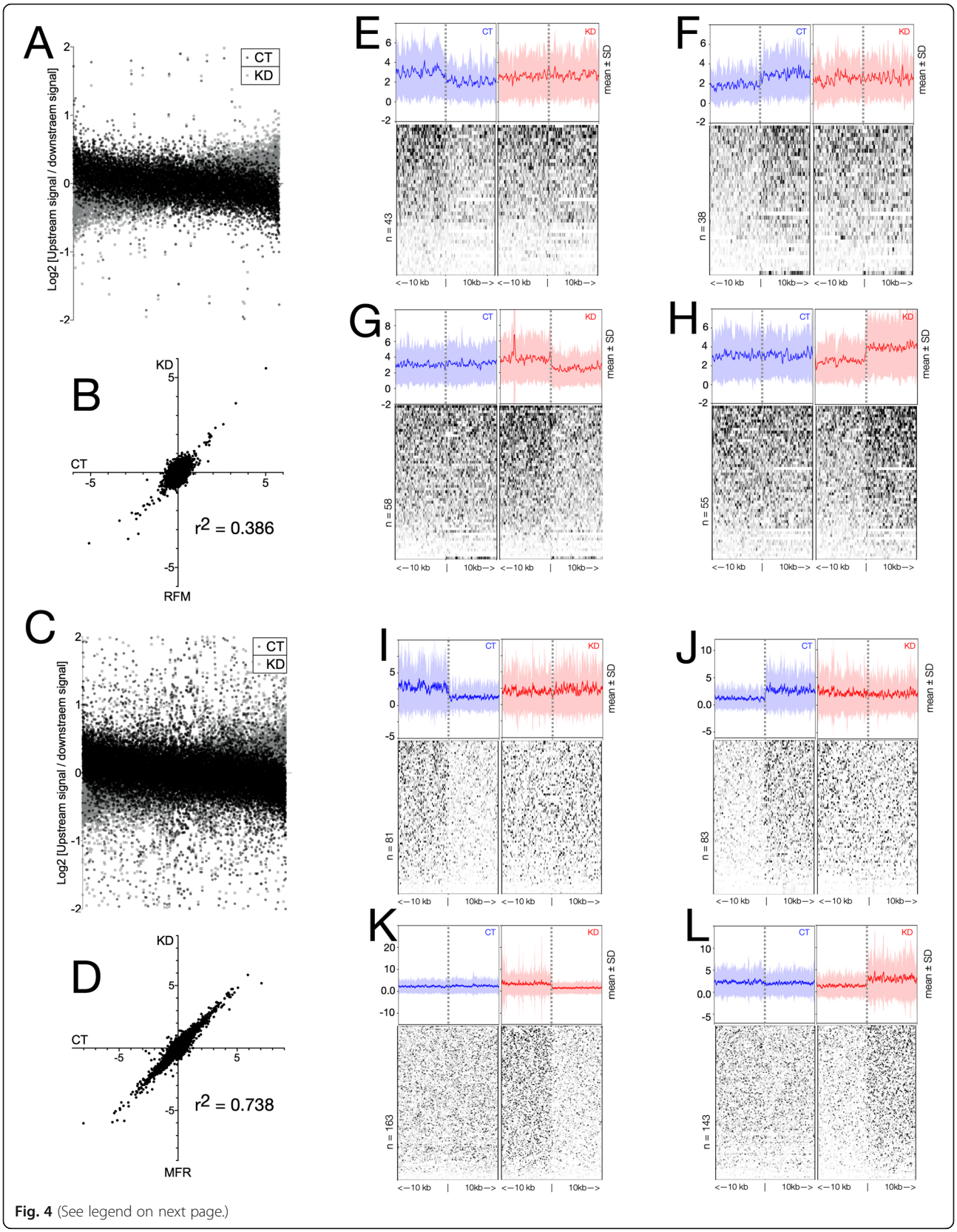

Fig. 4 (See legend on next page.)

(See figure on previous page.)
Fig. 4 CGGBP1-dependent regulation of cytosine methylation spread by RFM CTCF-binding sites. **a** to **d** Cytosine methylation signals in 10 kb flanks of RFM or MFR CTCF-binding sites suggest a barrier-like function for RFM. Upstream and downstream MeDIP-signals at RFM (**a**) show a weaker correlation (Pearson $r^2 = 0.386$) (**b**) than that at MFR (**c**), which show higher asymmetry with a better correlation (Pearson $r^2 = 0.738$) (**d**). **e** to **h** Cytosine methylation signal asymmetry at HEK293T RFM is lost due to an increase in cytosine methylation downstream (**e**) or upstream (**f**) of the CTCF-binding sites. Conversely, the asymmetry is gained due to a decrease in cytosine methylation downstream (**g**) or upstream (**h**). **i** to **l** The cytosine methylation asymmetry in RFM flanks in GM02639 is stronger than that observed in HEK293T and is lost due to an increase in methylation downstream (**i**) or (**j**). Similarly, a much stronger cytosine methylation asymmetry is achieved in GM02639 RFM than the HEK293T RFM due to a decrease in methylation downstream (**k**) or upstream (**l**). The stronger asymmetry of MeDIP signal in RFM flanks of GM02639 as compared to HEK293T reinforces lower stochasticity of methylation change caused by CGGBP1 depletion in the former

to Mat/– in KD (due to a random allelic choice for methylation; yellow shade in Fig. 5f). The loci falling in the purple, aqua and yellow shaded regions of the Fig. 5f represent regions at which cytosine methylation occurs with a stochastic allelic bias that depends on CGGBP1. The red-shaded regions represent a stochastic deviation from heterozygosity independent of CGGBP1 resulting in correlated parent-of-origin identities between CT and KD (Fig. 5f). To quantify the effect of CGGBP1 depletion on stochastic allelic choices and resulting deviations from heterozygosity, we calculated the differences in distribution of alleles with parental identities between CT and KD. As shown in Fig. 5g, the extreme allelic bias with a completely monoallelic methylation for the maternal as well as the paternal alleles was increased after CGGBP1 depletion. This was concomitant with a loss of biallelic methylation similarly for the maternal or the paternal alleles. The CTCF-binding sites were not associated with any specific type of allelic bias (data not shown). Moreover, we ensured that the allelic variation in methylation

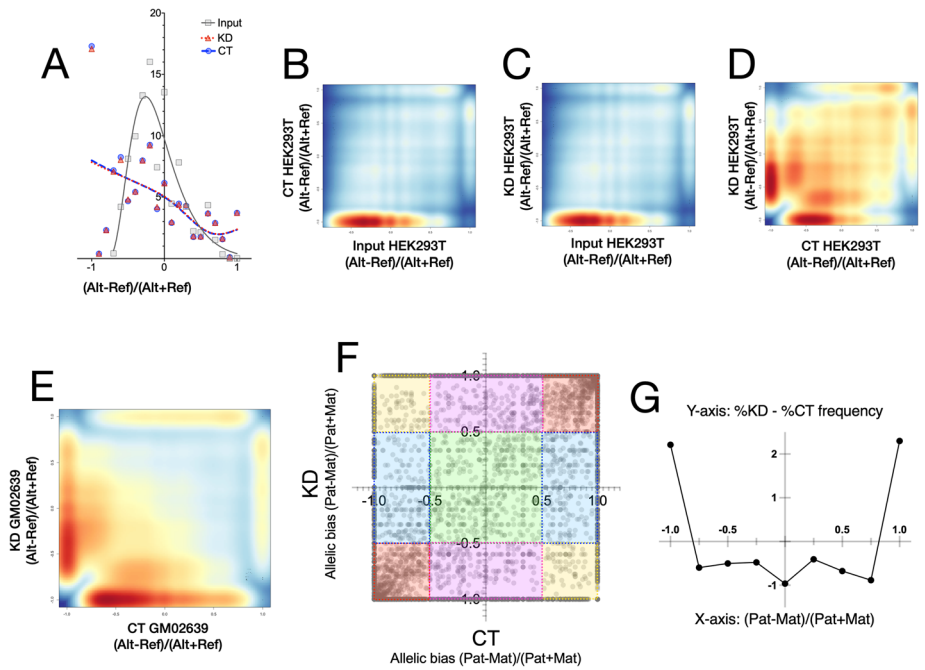

Fig. 5 Stochastic allelic imbalances in cytosine methylation are exacerbated by CGGBP1 depletion**a** Allelic imbalances in HEK293T MeDIP are strongly different from the allelic ratios expected in the input with negligible differences between CT and KD. **b** and **c** The poor correlation between allelic imbalances in the input and CT (**b**) or KD (**c**) is evident in the heat scatters. **d** The stochastic allelic choices for methylation give rise to a random distribution of allelic representations in HEK293T CT and KD MeDIP ($r^2 = 0.361$). **e** The allelic imbalances between GM02639 CT and KD were however highly anticorrelated as expected due to less stochasticity in these samples ($r^2 = 0.013$). **f** and **g** Parent-of-origin specific allelic imbalance scatter plot between GM02639 CT and KD (**f**) shows that CGGBP1 depletion enhances a random monoallelic bias with no preference for the maternal or the paternal alleles. For all locations where the CT and KD are expected to be heterozygous [(Pat-Mat)/(Pat+Mat)] values = 0 ± 0.5), the observed heterozygosity was very low (green shade; see text for details). Instead the unbiased monoallelic methylation with no difference between CT and KD ([(Pat-Mat)/(Pat+Mat)] values < − 0.5 or > 0.5) were observed for the majority of loci (red shades). A skewed allelic imbalance only in CT (aqua shade) or KD (pink shade). For a proportionate subset of loci, the parent-of-origin identities of the methylated alleles were reversed between CT and KD (yellow shades). **g** A frequency plot of the differences between percent of maternal and paternal allele contributions to the MeDIP DNA shows that CGGBP1 depletion minimizes allelic equivalence (X-axis = 0, Y-axis = − 1) and enhances exclusive maternal or paternal methylation (positive Y-axis values for X-axis = − 1 or + 1). Mat and Pat stand for Maternal and Paternal alleles respectively.

at CTCF-binding sites was not associated with CpG SNPs. The MeDIP-seq reads with allelic imbalances (biallelic-to-monoallelic as well as monoallelic-to-biallelic) were pooled and subjected to identification of CTCF-motifs EMBL_M1, EMBL_M2, REN_20, MIT_LM2, MIT_LM7 and MIT_LM23. The CpG counts of these motifs showed that CpG dinucleotides were absent in the CTCF-binding motifs showing AIM, even if they could be detected within the sequence reads containing those CTCF-binding sites (Additional file 15). We also compared these CGGBP1-dependent allelic imbalances in methylation with the allele-specific and haplotype-specific methylation reported by Do et al [33] and Bell et al [34] respectively. For every 1 million MeDIP-seq reads exhibiting a Ref-to-Alt or an Alt-to-Ref bias, less than 50 reads covered these sites with allele-specific methylation at imprinted sites and quantitative trait loci. Of the 7172 regions with haplotype-specific methylation regions, only 24 were represented in the MeDIP-seq regions with allelic imbalances. Thus, the allelic imbalances present in the CT and KD MeDIP-seq data are not concentrated in regions with non-stochastic allelic biases in methylation. These results confirmed that CGGBP1 depletion exacerbates the allelic imbalance in cytosine methylation in a stochastic manner giving rise to maternal or paternal biases in methylation as well as an allelic reversal of methylation. The cytosine methylation represented in the CT data itself has a considerable allelic imbalance and that the stochastic change in cytosine methylation caused by CGGBP1 depletion further enhances the extent of allelic imbalance.

Discussion

In this study, we have assayed the effects of CGGBP1 depletion on global patterns of cytosine methylation in two different cell types using MeDIP. Unlike WGBS, MeDIP provides a lower resolution information but offers an advantage in alignment frequency of sequence reads generating a higher effective coverage per locus than WGBS. MeDIP also allows querying of the methylation data for sequence patterns and motifs with a higher confidence than WGBS. Using WGBS in 1064Sk cells, we have previously reported that the methylation changes caused by CGGBP1 depletion are a near random combination of GoM and LoM [27, 28]. The deductions of gain or loss of methylation in a typical WGBS assay are confounded by two related reasons: (i) the WGBS technique captures and reports unmethylated as well as methylated cytosines without any weightage of the density of methylated or unmethylated cytosines [35], and (ii) the Bayesian probability frameworks in which cytosine methylation changes are called from a WGBS experiment rely on local cytosine and methylcytosine densities [36–39]. Thus, the large fraction of cytosines, that remains unmethylated or retains methylation between CT and KD fibroblasts in WGBS assays, confounds the distinction between deterministic or stochastic changes in methylation. In this study, to better understand the mechanisms of cytosine methylation regulation by CGGBP1, we employed MeDIP as a complement to our previous WGBS approaches [27, 28]. Unlike WGBS, MeDIP-seq sacrifices the base level resolution of methylation information for a semiquantitative estimation of methylation. The length resolution of methylation signal in MeDIP-seq is governed by the size of the input DNA fragments which is also reflected in the mean read lengths. The representation of a region in the MeDIP-seq thus depends on local methylcytosine density in a particular region and the frequency with which it is captured in MeDIP. These parameters are expected to vary intracellularly between alleles as well as due to intercellular heterogeneity in cytosine methylation. A thorough analysis of the MeDIP-seq data from CT and KD allows us to measure stochastic or deterministic quantitative changes in cytosine methylation over short sequences. In this case, we restricted our analyses to a minimum resolution of 0.2 kb (a convenient bin size that is larger than the mean length of the sequence reads). For a comparison, we have analyzed MeDIP-seq in CT and KD samples of fibroblasts (a cell type in which CGGBP1-regulated methylation has been studies earlier) and HEK293T (cells which are less sensitive to CGGBP1 depletion than fibroblasts). Moreover, in HEK293T cells we have recently shown that CGGBP1 determines the chromatin occupancy of CTCF at repeats versus motifs [20] and the current findings of methylation change in KD allow us to interpret cytosine methylation as a possible means through which CGGBP1 regulates CTCF occupancy. We have analyzed the nature of methylation changes in the two cell types and characterized the stochasticity of methylation changes caused by CGGBP1 depletion. We have also combined a blind TFBS analysis with our prior knowledge of the CTCF-CGGBP1 axis to extract subtle but deterministic methylation changes at CGGBP1-regulated the CTCF binding sites that are RFM.

The equal and mirroring patterns of GoM and LoM in HEK293T CT and KD are explainable as an outcome of stochastic changes in methylation. A wide range of methylation differences between HEK293T CT and KD corresponded to a similar high level of entropy. A non-functional TP53 in stem cells induces de novo methyltransferases resulting in high global methylation that is less prone to decrease after prolonged culturing [40, 41]. The SV40 T antigen-mediated inactivation of functional TP53 in HEK293T is thus expected to have a much higher buffering capacity against methylation changes. The relatively less dependence of HEK293T on CGGBP1 and the higher levels of methylation observed in it could thus be explained by stochastic nature of methylation changes caused by CGGBP1 depletion. The absence of

TP53 in HEK293T could further augment the stochasticity of methylation and dilute any deterministic changes. In the light of our recent findings that CGGBP1 is required for proper CTCF occupancy on the chromatin in HEK293T [20], the CGGBP1-regulated CTCF binding sites are strong candidate regions where a quantitative change could be expected between CT and KD. We reason that the high level of methylation stochasticity in HEK293T precludes the detection of methylation changes at CTCF binding sites. As a corollary, in fibroblasts, which are very sensitive to CGGBP1 depletion, do not have aberrant TP53-driven de novo methylation activity and may loose methylation further upon prolonged culturing, less entropy was observed in methylation changes. Supporting this, we observed that the stochasticity in fibroblasts was higher in regions with weaker methylation change. The regions with stronger change in methylation between CT and KD showed more deterministic change. The PCA accordingly revealed that RFM were the drivers of this deterministic change in methylation in fibroblasts. A higher GoM than LoM in GM02639 was different from the pattern seen in HEK293T. The directional change in cytosine methylation in GM02639 was detectable due to a lesser stochasticity than HEK293T, especially at higher levels of GoM. This deterministic change in methylation was concentrated at RFM. Interestingly, despite 90% knockdown of CGGBP1, HEK293T maintained strongly stochastic cytosine methylation in CT and KD, whereas just 50% knockdown in GM02639 caused a much less stochastic change in GM02639. This shows that there is a cell type specificity in the cytosine methylation changes caused by CGGBP1 knockdown and it is consistent with the previous findings [20] that HEK293T express higher levels of CGGBP1 and show less dependence on it as compared to fibroblasts. The stochasticity is observed in a large population sum total of methylation signals derived from pools of cells with mosaic methylation patterns.

Primary cells in culture rapidly lose methylation whereas immortalized cells are resistant to rapid changes in methylation under the same conditions. In our experiments we also see that the net quantitative change of methylation is close to zero in HEK293T cells whereas the fibroblasts show a much stronger net increase. The very high entropy of methylation patterns in HEK293T CT and KD compared to those in GM02639 are difficult to explain through random intercellular variations of methylation patterns as unlike WGBS data, we can not deconvolute [42] the MeDIP data to predict the cellular heterogeneity in the two cell cultures and any differences between them. Given that CGGBP1 depletion slows down or arrests cell cycle [43–45], the intercellular heterogeneity is expected to remain unaffected or diminish in KD as compared to CT. These facts suggest that a major fraction of stochasticity in methylation patterns, that is retained after CGGBP1 depletion, is due to factors other than intercellular heterogeneity. There is overwhelming evidence that the stochasticity of methylation patterns are due to localized allelic imbalances in methylation [3, 30, 46–48]. We tested the possibility that between CT and KD the stochastic changes in methylation patterns are due to random inter-allelic differences in methylation levels. Indeed we found that in both the cell types there were unexpectedly high levels of allelic imbalances in methylation. It was intriguing that between CT and KD, the allelic imbalances were qualitatively different. This included mostly a gain or loss of monoallelic or biallelic methylation. A significant subset however showed a highly unexpected monoallelic swap between the two parental genotypes in methylation between CT and KD. The allelic switch of methylation requires a stochasticity in methylation that is very dynamic and highly entropic. Since methylation as well as CTCF binding are known to be key regulators of genomic imprinting [8, 13, 49], we needed to rule out the possibility of a non-random allelic choice of methylation upon CGGBP1 depletion. To ensure that there was no parent of origin bias in the allelic switch between CT and KD, we sequenced and characterized the parental DNA of the fibroblasts. The allelic imbalance analysis with parent of origins defined convincingly demonstrated that the allelic imbalance of methylation in CT and KD are not biased towards any parent of origin and thus highly stochastic in nature.

Previously, WGBS has shown that the CpG methylation increases as well as decreases at repeats although the majority of methylation is at CHG and CHH cytosines [27]. Including all the three contexts, the prevalence of CHH cytosine methylation leads to the identification of G/C skew as a signature of sequences showing an increase as well as decrease due to CGGBP1 depletion [28]. The current MeDIP seq characterizes methylation patterns in a context independent manner and only much stronger differences in methylation (than those characterized through WGBS) would be able to affect the enrichment with methylcytosine antibody. Thus, even under high levels of stochasticity, the identification of quantitative differences in MeDIP signals is a strategy that has allowed us identification of a panel of TFBSs as targets. The strength of our TFBS strategy is a search for motifs that occur commonly in two disparate cell types used in this study. This ensured that the motif discovery was not due to cell type specific stochasticity in the methylation patterns but just due to the differences between CT and KD. This approach is prone to false negatives but robust against false positive motif detections. Binding of some TFs to their target sequences determines methylation turnover by regulating the

access of methylation machinery to the DNA [50]. This is known for CGGBP1 [25] and its depletion can thus lead to a random in methylation levels. In addition, methylation regulation by CTCF, REST, SP1, EGR1 has been described [50–53] and in a double blind search we have found these factors to be significantly enriched in the differentially methylated genomic bins between CT and KD. Thus the TFBS commonly identified in the differentially methylated bins of HEK293T as well as fibroblasts lead us to identify functionally relevant non-stochastic methylation changes caused by CGGBP1 depletion. The DNA-binding of TFs with binding site overrepresentation in differentially methylated bins are regulated by cytosine methylation as well as regulate cytosine methylation. Of all the TFs identified as significantly overrepresented in the CT KD DMRs, CTCF is the one for which a regulation by CGGBP1 has been demonstrated [20]. One of the proposed mechanisms of methylation regulation by CTCF, consistent with the stochasticity of methylation changes observed between CT and KD, is that CTCF sterically hinders methylation machinery access to the DNA [54]. We thus followed up the CGGBP1-regulated CTCF-binding sites to extract regions of non-stochastic changes. We have reported that CGGBP1 regulation of CTCF occupancy at repeats and motifs are inversely related [20]. There is no evidence that CTCF-repeat interaction is indirect and thus different from the direct binding of CTCF to its motif. Thus, the methylation sensitivity of CTCF binding at motifs may not necessarily apply to CTCF binding at repeats. This is supported by the findings that L1 repeats are not differentially represented in differentially methylated regions, but the CTCF motifs are specifically differentially methylated between CT and KD. Following up on this lead, we segregated the CTCF binding sites as RFM and MFR and confirmed that CGGBP1 regulated methylation changes affect CTCF-binding motifs and not the repeats non-stochastically. Our findings suggest that different cell types show different types of methylation changes at CTCF motifs upon CGGBP1 depletion. These results also suggest that motif-specific methylation change may be a mechanism underlying the shift in CTCF binding preferences from repeats to motifs upon CGGBP1 depletion.

Stochasticity is an innate property of cytosine methylation and has been addressed in multiple studies [30, 55, 56]. In our investigation here, the stochastic nature of methylation has remained dominant over the effects of CGGBP1 depletion on global methylation patterns. The fundamental question of why CGGBP1 depletion leads to such a widespread resetting of methylation remains unknown. We postulate that the widespread occupancy of CGGBP1 on the genome in the presence of normal amounts of CGGBP1 maintains a state of dynamic equillibrium wherein the CGGBP1-bound DNA remains unavailable for binding and activity of the methylation regulatory apparatus. The lowering of CGGBP1 levels disrupts this equilibrium such that the DNA denuded of CGGBP1 becomes more amenable to activity by the methylation regulatory apparatus.

Our results suggest an interplay between a stochastic disruption in methylation caused by CGGBP1 depletion and its effects on specific TFBSs, including CTCF. The disruption of methylation upon CGGBP1 depletion targets sites at which CTCF binding has been recently demonstrated. The cause-consequence relationship between methylation changes and CTCF binding at RFM remains unknown, but it is highly likely that it is a two-way feedback process. However, the disruption in CTCF occupancy at RFM seems to be functionally relevant as the methylation asymmetry in the flanks of these CTCF-binding sites seem to be affected specifically. These results complement our previous findings that the H3K9me3 signals exhibit asymmetry in the flanks on CGGBP1-regulated CTCF-binding repeats. It appears that CGGBP1 stabilizes a cytosine methylation profile at RFM that allows CTCF to maintain a barrier against methylation spread across the RFM. Thus, cytosine methylation homeostasis is a crucial entity at the interface of regulation of CTCF barrier activities by CGGBP1.

Conclusion

CGGBP1 depletion induced cytosine methylation changes are stochastic in HEK293T and GM02639 cells. CGGBP1 depletion in these cells changes cytosine methylation patterns such that the stochasticity remains unperturbed but the allelic imbalances that underlie the stochasticity are different in CT and KD. Embedded in the largely stochastic methylation patterns in CT and KD are specific TFBSs which show a cell type-specific quantitative change in methylation. One of these TFs is CTCF. The methylation patterns at CTCF-binding MFR and RFM show different dependence on CGGBP1. The MFR methylation remains stochastic between CT and KD whereas the RFM shows a non-stochastic change. The methylation changes between CT and KD were stochastic with no parent-of-origin bias. The non-stochastic methylation changes at RFM were not due to lower levels of stochastic allelic imbalances.

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s12863-020-00894-8.

Additional file 1. Primer names, locations, sequences and annealing temperatures used for the candidate region methylation analyses shown in Additional file 14. The cyclic denaturation (95 °C, 30 s) and extension

(70 °C) steps were the same for all primer combinations. Data was collected post-extension at 80 °C as mentioned in methods.

Additional file 2. CGGBP1 depletion in HEK293T cells: HEK293T cells were transduced with non-targeting shRNA or CGGBP1-targeting shRNA lentiviruses. Lentivirus-transduced cells were selected with 10 μg/ml Puromycin for 1 week and subjected to immunoblotting. The level of CGGBP1 and GAPDH are shown in the upper and lower panel respectively. A CGGBP1 knockdown of approximately 77% is observed when normalized to GAPDH levels.

Additional file 3. Tabulation of the sequencing and alignment statistics for CT and KD MeDIP in HEK293T and GM02639.

Additional file 4. Table represents GC content and percentage enrichment of CpG and non-CpG context cytosines in MeDIP read sequences for CT and KD in HEK293T and GM02639. The percentage of CpG, CHG and CHH represents the relative abundance for each cytosine context as a percentage of the total methylated cytosines enriched in MeDIP for CT and KD in HEK293T and GM02639.

Additional file 5. Methylation differences between CT and KD are discernible at small genomic length ranges. Genome-wide methylation signal distribution was compared between CT and KD by using "deeptools multiBigwigSummary". Methylation signals were compared at bin sizes of 10 kb, 5 kb, 1 kb and 0.2 kb. Correlation between CT and KD was computed by the Spearman method by using "deeptools plotCorrelation" for HEK293T cells.

Additional file 6. The MeDIP signal correlation between CT and KD at different genomic bin sizes decline with a reduction in bin size specifically as randomization of coordinates (and thus corresponding sequences) changes the correlation stochastically away from the observed correlation coefficients with actual MeDIP sequences (refer to Additional file 5 for a comparison with correlation coefficients without any randomization).

Additional file 7. The MeDIP reads distribution for HEK293T CT and KD at low methylation signal bins (1 to 4): Lower methylation signal bins account for a major fraction of MeDIP reads. The frequency of the MeDIP reads for these bins with less than 5 methylation read signals were calculated separately from those in the range 5–30.

Additional file 8. CGGBP1 depletion in GM02639 cells: GM02639 cells were transfected with non-targeting or CGGBP1-targeting siRNA twice at 24 and 72 h post-seeding. Cells were harvested at 96 h. Immunoblotting results for CGGBP1 (upper panel) and GAPDH (lower panel) show approximately 55% knockdown of CGGBP1 when normalized to the level of GAPDH.

Additional file 9. Methylation differences between CT and KD are discernible at small genomic length ranges: Genome-wide methylation signal distribution was compared between CT and KD by using "deeptools multiBigwigSummary". Methylation signals were compared at bin sizes of 10 kb, 5 kb, 1 kb and 0.2 kb. Correlation between CT and KD was computed by the Spearman method by using "deeptools plotCorrelation" for GM02639 cells. These correlation coefficients can be compared with those for HEK293T (Additional file 7).

Additional file 10. The figure shows the methylation reads distribution in CT and KD at lower methylation bins (1 to 4) in GM02639 cells.

Additional file 11. Repeat content analysis in HEK293T CT and KD MeDIP DNA shows subfamily-specific methylation changes: Methylation bin frequency plots for HEK293T CT and KD. CT and KD reads for each methylation bin (From 5 to 30) were merged and sequences for merged regions (> 150 bp long) were extracted and subjected to repeat identification. The figures depict the occurrence of the three most populous repeats (Satellites, L1-LINEs and Alu-SINEs). (A) No overall differences in repeat content were observed between CT and KD. (B) A classification of the Alu SINEs into J, S and Y subfamilies revealed subfamily-specific differences in methylation between CT and KD. AluJb and AluSx showed consistently higher methylation in KD across all the methylation bins (top two panels), while AluY showed lower methylation in KD (bottom panel). (C) A subfamily classification of L1 repeats revealed that the L1HS and P family LINE1 such as L1P1, L1P2 and L1PA4 are reduced in KD although these repeat subtypes are more prevalent in highly methylated regions. (D) In contrast, the early originated LINE1 such as L1M1, L1M3, L1M4 and L1M5 showed increased methylation in KD and these repeat subtypes are prevalent in regions with low levels of methylation.

Additional file 12. Repeat content analysis in GM02639 CT and KD MeDIP DNA shows subtle subfamily-specific methylation changes. CT and KD MeDIP repeat identification was performed as described for data in the Additional file 7. repeat identification and the occurrence of the three most populous repeats (Satellites, L1-LINEs and Alu-SINEs) were analyzed. (A) Satellite and L1 repeats were overrepresented and underrepresented respectively in KD. No difference in Alu-SINEs content was observed however. (B) Unlike HEK293T data, the subfamily classification of Alus does not reveal any subfamily-specific methylation differences between CT and KD in GM02639. (C) and (D): L1 repeat subfamily classification showed no consistent differences in methylation between CT and KD.

Additional file 13. Highly similar CTCF binding motifs are present in regions undergoing GoM, LoM or showing no methylation change upon CGGBP1 depletion. Methylation signals for CT and KD were calculated for each 0.2 kb bin for HEK293T and bins were grouped into GoM, LoM and "No change" (as described in method in details). CTCF motif positive 0.2 kb bins were filtered out from each group. Motif positive GoM, LoM and "No change" 0.2 kb bin sequences were subjected to de novo motif search by using MEME suite.

Additional file 14 Quantitative PCR (double delta analysis of relative changes in levels of methylated DNA) on CT and KD DNA from human dermal fibroblasts shows widespread differences between CT and KD. (A) qPCR on *Hpa*II-digested DNA (digests DNA flanking unmethylated cytosine) shows a significant ($p < 0.05$, $n = 3$ technical replicates, unpaired T test) gain of methylation at CpG sites for multiple genomic regions representing the PEG10 locus, GRB10 locus and a CpG island termed CpG-16. Conversely, a loci representing TDG, CTCF-binding site termed CTCF-2-2, TET3 and NAP1L5 showed a loss of methylation. All *Hpa*II Ct values were normalized against corresponding Ct values obtained after *Msp*I digestion. (B) Widespread CpG methylation disturbances (with no significance) were also observed at multiple other loci. (C) qPCRs on *Mcr*BC-digested CT and KD DNA showed a loss of methylation at non-CpG sites for multiple loci including Alu repeats and CTCF-binding sites ($p < 0.05$, $n = 3$ technical replicates, unpaired T test). (D) Several locations displayed widespread methylation disturbances (with no significance) as revealed by *Mcr*BC-digestion. Refer to Additional file 1 for exact location and PCR details.

Additional file 15. CpG SNPs in the entire AIM dataset and those mapping to the CTCF motifs undergoing AIM in GM02639 cells. The methylation changes at CTCF motifs are not linked to any detectable CpG SNPs within the motifs even if the flanking regions in the sequence reads harbour CpG SNPs.

Abbreviations

CT: Non-targeting shRNA or siRNA sample; KD: CGGBP1-targeting shRNA or siRNA sample; RFM: Repeat-free motifs; MFR: Motif-free repeats; LoM: Loss of methylation; GoM: Gain of methylation; MeDIP-seq: Methyl(cytosine) DNA immunoprecipitation sequencing; Diff/Sum: Difference/Sum ratio

Acknowledgements

The authors acknowledge the Gujarat Biotechnology Research Centre for sequencing services, Coriell cell repository for fibroblasts (GM02640, GM02641 and GM02639), NCCS Pune for HEK293T cells and Mr. Sudeep N Banerjee (ISTF, IITGN) for help with computational resources.

Authors' contributions

MP, DP and SD performed the experiments. All authors participated in data analysis and manuscript writing. US supervised all aspects of the work. The manuscript is read and approved by all the authors.

Funding

Grants to US from Gujarat State Biotechnology Mission FAP-1337 (SSA/4873), SERB EMR/2015/001080, DST-ICPS T-357, DBT BT/PR15883/BRB/10/1480/2016 and Biomedical Engineering Centre IITGN, Indian Institute of Technology Gandhinagar. The studentships of MP and DP were supported by UGC-NET JRF and SD from MHRD, GoI and DBT BT/PR15883/BRB/10/1480/2016.

Availability of data and materials
The datasets generated in this study are publicly accessible at NCBI GEO vide ID GSE145300.

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Received: 26 March 2020 Accepted: 23 July 2020
Published online: 29 July 2020

References
1. Ong C-T, Corces VG. CTCF: an architectural protein bridging genome topology and function. Nat Rev Genet. 2014;15:234–46.
2. Cuddapah S, Jothi R, Schones DE, Roh T-Y, Cui K, Zhao K. Global analysis of the insulator binding protein CTCF in chromatin barrier regions reveals demarcation of active and repressive domains. Genome Res. 2009;19:24–32.
3. Tang Z, Luo OJ, Li X, Zheng M, Zhu JJ, Szalaj P, et al. CTCF-mediated human 3D genome architecture reveals chromatin topology for transcription. Cell. 2015;163:1611–27.
4. Schmidt D, Schwalie PC, Wilson MD, Ballester B, Gonçalves A, Kutter C, et al. Waves of retrotransposon expansion remodel genome organization and CTCF binding in multiple mammalian lineages. Cell. 2012;148:335–48.
5. Kim TH, Abdullaev ZK, Smith AD, Ching KA, Loukinov DI, Green RD, et al. Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome. Cell. 2007;128:1231–45.
6. Xie X, Mikkelsen TS, Gnirke A, Lindblad-Toh K, Kellis M, Lander ES. Systematic discovery of regulatory motifs in conserved regions of the human genome, including thousands of CTCF insulator sites. Proc Natl Acad Sci U S A. 2007; 104:7145–50.
7. Hashimoto H, Wang D, Horton JR, Zhang X, Corces VG, Cheng X. Structural Basis for the Versatile and Methylation-Dependent Binding of CTCF to DNA. Mol Cell. 2017;66:711–20 e3.
8. Bell AC, Felsenfeld G. Methylation of a CTCF-dependent boundary controls imprinted expression of the Igf2 gene. Nature. 2000;405:482–5.
9. Filippova GN, Thienes CP, Penn BH, Cho DH, Hu YJ, Moore JM, et al. CTCF-binding sites flank CTG/CAG repeats and form a methylation-sensitive insulator at the DM1 locus. Nat Genet. 2001;28:335–43. https://doi.org/10.1038/ng570.
10. Szabó P, Tang SH, Rentsendorj A, Pfeifer GP, Mann JR. Maternal-specific footprints at putative CTCF sites in the H19 imprinting control region give evidence for insulator function. Curr Biol. 2000;10:607–10.
11. Kanduri M, Kanduri C, Mariano P, Vostrov AA, Quitschke W, Lobanenkov V, et al. Multiple nucleosome positioning sites regulate the CTCF-mediated insulator function of the H19 imprinting control region. Mol Cell Biol. 2002; 22:3339–44. https://doi.org/10.1128/mcb.22.10.3339-3344.2002.
12. Pant V. The nucleotides responsible for the direct physical contact between the chromatin insulator protein CTCF and the H19 imprinting control region manifest parent of origin-specific long-distance insulation and methylation-free domains. Genes Dev. 2003;17:586–90. https://doi.org/10.1101/gad.254903.
13. Adalsteinsson B, Ferguson-Smith A. Epigenetic control of the genome—lessons from genomic imprinting. Genes. 2014;5:635–55. https://doi.org/10.3390/genes5030635.
14. Herold M, Bartkuhn M, Renkawitz R. CTCF: insights into insulator function during development. Development. 2012;139:1045–57.
15. Weintraub AS, Li CH, Zamudio AV, Sigova AA, Hannett NM, Day DS, et al. YY1 Is a Structural Regulator of Enhancer-Promoter Loops. Cell. 2017;171: 1573–88 e28.
16. Li Y, Huang W, Niu L, Umbach DM, Covo S, Li L. Characterization of constitutive CTCF/cohesin loci: a possible role in establishing topological domains in mammalian genomes. BMC Genomics. 2013;14:553.
17. Handoko L, Xu H, Li G, Ngan CY, Chew E, Schnapp M, et al. CTCF-mediated functional chromatin interactome in pluripotent cells. Nat Genet. 2011;43: 630–8.
18. Hou C, Dale R, Dean A. Cell type specificity of chromatin organization mediated by CTCF and cohesin. Proc Natl Acad Sci U S A. 2010;107:3651–6.
19. Hsu SC, Gilgenast TG, Bartman CR, Edwards CR, Stonestrom AJ, Huang P, et al. The BET Protein BRD2 Cooperates with CTCF to Enforce Transcriptional and Architectural Boundaries. Molecular Cell. 2017;66:102–16.e7. https://doi.org/10.1016/j.molcel.2017.02.027.
20. Patel D, Patel M, Datta S, Singh U. CGGBP1 regulates CTCF occupancy at repeats. Epigenetics Chromatin. 2019;12:57.
21. Naumann F, Remus R, Schmitz B, Doerfler W. Gene structure and expression of the 5′-(CGG)n-3′-binding protein (CGGBP1). Genomics. 2004;83:106–18. https://doi.org/10.1016/s0888-7543(03)00212-x.
22. Naumann A, Kraus C, Hoogeveen A, Ramirez CM, Doerfler W. Stable DNA methylation boundaries and expanded trinucleotide repeats: role of DNA insertions. J Mol Biol. 2014;426:2554–66.
23. Müller-Hartmann H, Deissler H, Naumann F, Schmitz B, Schröer J, Doerfler W. The human 20-kDa 5′-(CGG)n-3′-binding protein is targeted to the nucleus and affects the activity of theFMR1Promoter. J Biol Chem. 2000;275: 6447–52. https://doi.org/10.1074/jbc.275.9.6447.
24. Deissler H, Wilm M, Genç B, Schmitz B, Ternes T, Naumann F, et al. Rapid protein sequencing by tandem mass spectrometry and cDNA cloning of p20-CGGBP. J Biol Chem. 1997;272:16761–8. https://doi.org/10.1074/jbc.272.27.16761.
25. Deissler H, Behn-Krappa A, Doerfler W. Purification of nuclear proteins from human HeLa cells that bind specifically to the unstable tandem repeat (CGG) in the human FMR1 gene. J Biol Chem. 1996;271:4327–34. https://doi.org/10.1074/jbc.271.8.4327.
26. Goracci M, Lanni S, Mancano G, Palumbo F, Chiurazzi P, Neri G, et al. Defining the role of the CGGBP1 protein in FMR1 gene expression. Eur J Hum Genet. 2016;24:697–703.
27. Agarwal P, Collier P, Fritz MH-Y, Benes V, Wiklund HJ, Westermark B, et al. CGGBP1 mitigates cytosine methylation at repetitive DNA sequences. BMC Genomics. 2015;16:390.
28. Patel D, Patel M, Westermark B, Singh U. Dynamic bimodal changes in CpG and non-CpG methylation genome-wide upon CGGBP1 loss-of-function. BMC Res Notes. 2018;11:419.
29. Agarwal P, Enroth S, Teichmann M, Jernberg Wiklund H, Smit A, Westermark B, et al. Growth signals employ CGGBP1 to suppress transcription of Alu-SINEs. Cell Cycle. 2016;15:1558–71.
30. Onuchic V, Lurie E, Carrero I, Pawliczek P, Patel RY, Rozowsky J, et al. Allele-specific epigenome maps reveal sequence-dependent stochastic switching at regulatory loci. Science. 2018;361. https://doi.org/10.1126/science.aar3146.
31. Yin Y, Morgunova E, Jolma A, Kaasinen E, Sahu B, Khund-Sayeed S, et al. Impact of cytosine methylation on DNA binding specificities of human transcription factors. Science. 2017;356. https://doi.org/10.1126/science.aaj2239.
32. Clouaire T, Stancheva I. Methyl-CpG binding proteins: specialized transcriptional repressors or structural components of chromatin? Cell Mol Life Sci. 2008;65:1509–22. https://doi.org/10.1007/s00018-008-7324-y.
33. Do C, Lang CF, Lin J, Darbary H, Krupska I, Gaba A, et al. Mechanisms and disease associations of haplotype-dependent allele-specific DNA methylation. Am J Hum Genet. 2016;98:934–55.
34. Bell CG, Gao F, Yuan W, Roos L, Acton RJ, Xia Y, et al. Obligatory and facilitative allelic variation in the DNA methylome within common disease-associated loci. Nat Commun. 2018;9:8.
35. Yong W-S, Hsu F-M, Chen P-Y. Profiling genome-wide DNA methylation. Epigenetics Chromatin. 2016;9. https://doi.org/10.1186/s13072-016-0075-3.
36. Feng H, Conneely KN, Wu H. A Bayesian hierarchical model to detect differentially methylated loci from single nucleotide resolution sequencing data. Nucleic Acids Res. 2014;42:e69. https://doi.org/10.1093/nar/gku154.
37. Huh I, Yang X, Park T, Yi SV. Bis-class: a new classification tool of methylation status using bayes classifier and local methylation information. BMC Genomics. 2014;15:608.
38. Lock EF, Dunson DB. Bayesian genome- and epigenome-wide association studies with gene level dependence. Biometrics. 2017;73:1018–28.
39. Wang H, He C, Kushwaha G, Xu D, Qiu J. A full Bayesian partition model for identifying hypo- and hyper-methylated loci from single nucleotide resolution sequencing data. BMC Bioinformatics. 2016;17(Suppl 1):7.
40. Tovy A, Spiro A, McCarthy R, Shipony Z, Aylon Y, Allton K, et al. p53 is essential for DNA methylation homeostasis in naïve embryonic stem cells, and its loss promotes clonal heterogeneity. Genes Dev. 2017;31:959–72.
41. Wilson VL, Jones PA. DNA methylation decreases in aging but not in immortal cells. Science. 1983;220:1055–7.

42. Titus AJ, Gallimore RM, Salas LA, Christensen BC. Cell-type deconvolution from DNA methylation: a review of recent applications. Hum Mol Genet. 2017;26:R216–24.
43. Singh U, Maturi V, Jones RE, Paulsson Y, Baird DM, Westermark B. CGGBP1 phosphorylation constitutes a telomere-protection signal. Cell Cycle. 2014; 13:96–105.
44. Singh U, Roswall P, Uhrbom L, Westermark B. CGGBP1 regulates cell cycle in cancer cells. BMC Mol Biol. 2011;12:28.
45. Singh U, Westermark B. CGGBP1 is a nuclear and midbody protein regulating abscission. Exp Cell Res. 2011;317:143–50.
46. Jeffries AR, Perfect LW, Ledderose J, Schalkwyk LC, Bray NJ, Mill J, et al. Stochastic choice of allelic expression in human neural stem cells. Stem Cells. 2012;30:1938–47.
47. Song Y, van den Berg PR, Markoulaki S, Soldner F, Dall'Agnese A, Henninger JE, et al. Dynamic Enhancer DNA Methylation as Basis for Transcriptional and Cellular Heterogeneity of ESCs. Mol Cell. 2019;75:905–20 e6.
48. Luo Y, He J, Xu X, Sun M-A, Wu X, Lu X, et al. Integrative single-cell omics analyses reveal epigenetic heterogeneity in mouse embryonic stem cells. PLoS Comput Biol. 2018;14:e1006034.
49. Tost J, Jammes H, Dupont J-M, Buffat C, Robert B, Mignot T-M, et al. Non-random, individual-specific methylation profiles are present at the sixth CTCF binding site in the human H19/IGF2 imprinting control region. Nucleic Acids Res. 2006;34:5438–48.
50. Feldmann A, Ivanek R, Murr R, Gaidatzis D, Burger L, Schübeler D. Transcription factor occupancy can mediate active turnover of DNA methylation at regulatory regions. PLoS Genet. 2013;9:e1003994.
51. Kemp CJ, Moore JM, Moser R, Bernard B, Teater M, Smith LE, et al. CTCF haploinsufficiency destabilizes DNA methylation and predisposes to cancer. Cell Rep. 2014;7:1020–9.
52. Boumber YA, Kondo Y, Chen X, Shen L, Guo Y, Tellez C, et al. An Sp1/Sp3 binding polymorphism confers methylation protection. PLoS Genet. 2008;4: e1000162.
53. Sun Z, Xu X, He J, Murray A, Sun M-A, Wei X, et al. EGR1 recruits TET1 to shape the brain methylome during development and upon neuronal activity. Nat Commun. 2019;10:3892.
54. Zampieri M, Guastafierro T, Calabrese R, Ciccarone F, Bacalini MG, Reale A, et al. ADP-ribose polymers localized on Ctcf-Parp1-Dnmt1 complex prevent methylation of Ctcf target sites. Biochem J. 2012;441:645–52.
55. Reinius B, Sandberg R. Random monoallelic expression of autosomal genes: stochastic transcription and allele-level regulation. Nat Rev Genet. 2015;16: 653–64.
56. Feinberg AP, Irizarry RA. Evolution in health and medicine Sackler colloquium: stochastic epigenetic variation as a driving force of development, evolutionary adaptation, and disease. Proc Natl Acad Sci U S A. 2010;107(Suppl 1):1757–64.

Publisher's Note